Multiclass Learning with Margin: Exponential Rates with No Bias-Variance Trade-Off

Stefano Vigogna [1] **Giacomo Meanti** [2] **Ernesto De Vito** [3] **Lorenzo Rosasco** [2 4 5]

Abstract

We study the behavior of error bounds for multiclass classification under suitable margin conditions. For a wide variety of methods we prove that the classification error under a hard-margin condition decreases exponentially fast without any bias-variance trade-off. Different convergence rates can be obtained in correspondence of different margin assumptions. With a self-contained and instructive analysis we are able to generalize known results from the binary to the multiclass setting.

1. Introduction

It was recently remarked that the learning curves observed in practice can be quite different from those predicted in theory (Zhang et al., 2021). In particular, while one might expect performance to degrade as models get larger or less constrained (Hastie et al., 2009), this is in fact not the case. By the no free lunch theorem (Wolpert, 1996), theoretical results critically depend on the set of assumptions made on the problem. Such assumptions can be hard to verify in practice, hence a possible way to tackle the seeming contradictions in learning theory *vs.* practice is to consider a wider range of assumptions, and check whether the corresponding results can explain empirical observations.

In the context of classification, it is interesting to consider assumptions describing the difficulty of the problem in terms of *margin* (Mammen & Tsybakov, 1999; Tsybakov, 2004). It is well known that very different learning curves can be obtained depending on the considered margin condition (Bartlett et al., 2006). Further, the behavior of the test error in terms of misclassification can be considerably different from that induced by the surrogate loss function used for empirical risk minimization (Zhang, 2004a; Bartlett et al., 2006). An extreme case is when there is a hard margin among the classes. Indeed, in this case the misclassification error can decrease *exponentially* fast as the number of points increases, while the surrogate loss error displays a polynomial decay. This behavior was first noted in (Koltchinskii & Beznosova, 2005; Audibert & Tsybakov, 2007) for a wide class of estimators (see also (Yao et al., 2007)), and reprised more recently in (Pillaud-Vivien et al., 2018; Nitanda & Suzuki, 2019) for stochastic gradient descent. The effect of margin conditions has also been considered for *multiclass* learning (Zhang, 2004b; Chen & Sun, 2006; Mroueh et al., 2012), but not in the hard-margin case. Interestingly, hard-margin and exponential rates have been studied by (Cabannes et al., 2021) in the context of structured prediction (Nowak et al., 2019), a general framework which includes traditional classification. However, these latter results are restricted to least-squares-based estimators.

The purpose of our paper is twofold. On the one hand, we analyze the effect of margin conditions, and in particular hard-margin conditions, for a wide class of multiclass estimators derived from different surrogate losses. On the other hand, we build on ideas in (Mroueh et al., 2012; Pillaud-Vivien et al., 2018; Nitanda & Suzuki, 2019) to provide a simplified and self-contained treatment that naturally recovers results for binary classification as a special case. In particular, we note that, in the presence of a hard margin, the misclassification error curve does not exhibit any *bias-variance* trade-off, thus providing a possible explanation to the empirical observations that motivate our study.

The rest of the paper is organized as follows. We conclude the introduction by setting up some basic notation. In Section 2 we describe the multiclass classification problem, the surrogate approach and the simplex encoding. In Section 3 we analyze the bias-variance decomposition for the misclassification risk, discuss soft and hard-margin conditions, and prove our main results of exponential convergence under assumptions of hard margin. In Section 4 we validate the theory with experiments on synthetic data. Some final remarks are provided in Section 5.

[1]RoMaDS, University of Rome Tor Vergata, Rome, Italy [2]MaLGa - DIBRIS, University of Genova, Italy [3]MaLGa - DIMA, University of Genova, Italy [4]Istituto Italiano di Tecnologia, Genova, Italy [5]CBMM - MIT, Cambridge, MA, USA. Correspondence to: Stefano Vigogna <vigogna@mat.uniroma2.it>.

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

Notation. We will be using the following general notation. $a \lesssim b$ means that $a \leq cb$ for some positive absolute constant c. The Euclidean norm and inner product of vectors $w, w' \in \mathbb{R}^p$ are denoted by $\|w\|$ and $\langle w, w' \rangle$, respectively, while the outer product is denoted by $w \otimes w'$. For an event E, $\mathbb{1}\{E\}$ denotes its indicator function, and $\mathbb{P}\{E\}$ its probability. The expectation of a random variable Z is denoted by $\mathbb{E}Z$; when the expectation is taken only with respect to a random variable X (but Z possibly depends also on other variables), we write $\mathbb{E}_X Z$. Conditioning of events or random variables on an event E is indicated by $\cdot \mid E$. $L^0(\mathcal{X}, \mathcal{Z})$ is the space of measurable functions on the probability space $\mathcal{X}$ and with values in $\mathcal{Z} \subset \mathbb{R}^p$, and $L^\infty(\mathcal{X}, \mathcal{Z})$ the subspace of almost surely bounded functions, with norm $\| \cdot \|_\infty$.

2. Setting

We consider a standard multiclass learning problem. Let $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ be a random pair, where $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y}$ is a finite set of $T \geq 2$ elements. We call the elements of $\mathcal{Y}$ *classes*, and a (measurable) function $c : \mathcal{X} \to \mathcal{Y}$ a *classifier*. The *misclassification risk* of a classifier c is

$$\mathcal{R}(c) = \mathbb{P}\{c(X) \neq Y\}.$$

Let

$$\rho(y \mid x) = \mathbb{P}\{Y = y \mid X = x\}$$

denote the conditional probability of the class y given the observation x. The risk $\mathcal{R}$ is minimized by the *Bayes rule*

$$c_*(x) = \arg\max_{y \in \mathcal{Y}} \rho(y \mid x).$$

We denote the minimum risk by $\mathcal{R}_* = \mathcal{R}(c_*)$. Given n independent copies (X_i, Y_i) of (X, Y), $i = 1, \ldots, n$, the goal is to learn a classifier $\widehat{c}$ such that $\mathcal{R}(\widehat{c}) - \mathcal{R}_* \to 0$ in expectation as $n \to \infty$. More precisely, we are interested in finite-sample bounds of the form

$$\mathbb{E}\mathcal{R}(\widehat{c}) - \mathcal{R}_* \lesssim a_n,$$

where $a_n \to 0$ gives a rate of convergence.

Empirical risk minimization would prescribe to compute $\widehat{c}$ by minimizing a sample version of $\mathcal{R}$. The misclassification risk can be seen as the expectation of the *0-1 loss*

$$\mathbb{1}\{y \neq y'\}, \qquad y, y' \in \mathcal{Y}.$$

The empirical mean would thus be $\frac{1}{n} \sum_{i=1}^n \mathbb{1}\{c(X_i) \neq Y_i\}$. However, the 0-1 loss is neither smooth nor convex, and optimizing it is in general an NP-hard combinatorial problem (Feldman et al., 2012). A viable strategy is to replace the 0-1 loss with a convex *surrogate*, and the space of classifiers with a suitable linear space of vector-valued functions. To do this, it is necessary to choose a vector encoding of the classes $\mathcal{Y} \hookrightarrow \mathbb{R}^p$ and a decoding operator $D : \mathbb{R}^p \to \mathcal{Y}$. Following (Mroueh et al., 2012), we encode the T classes as the vertices of a $(T-1)$-simplex embedded in $\mathbb{R}^{T-1}$ (see Figure 1).

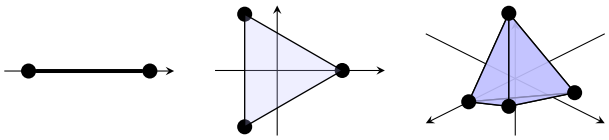


Figure 1. Simplex encoding for $T = 2, 3, 4$.

For notational convenience, we identify $\mathcal{Y}$ itself with its simplex encoding, that is, $\mathcal{Y}$ is the set of points in $\mathbb{R}^{T-1}$ such that

$$\|y\| = 1, \qquad \langle y, y' \rangle = -\tfrac{1}{T-1}.$$

The decoding operator assigns a vector to the class with largest projection, with ties arbitrarily broken (see Figure 2):

$$D : \mathbb{R}^{T-1} \to \mathcal{Y}, \qquad D(w) = \arg\max_{y \in \mathcal{Y}} \langle w, y \rangle.$$

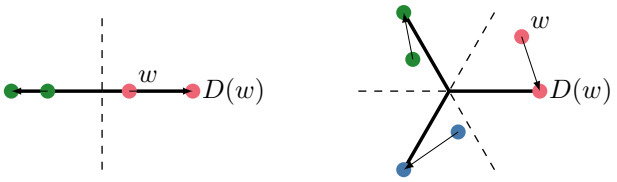


Figure 2. Simplex decoding ($T = 2, 3$).

In the case of binary classification ($T = 2$), we have $\mathcal{Y} = \{\pm 1\} \subset \mathbb{R}$ and $D(w) = \operatorname{sign}(w)$. A *plug-in* classifier $Df(x) = D(f(x))$ can be defined by composing a vector-valued function $f : \mathcal{X} \to \mathbb{R}^{T-1}$ with the decoding operator. The simplex coding offers some advantages over other common types of coding, such as one-hot. First, as we just saw, it is perfectly consistent with the standard $(\{\pm 1\}, \operatorname{sign})$ coding of binary classification. Second, it automatically satisfies structural constraints that other codings need to impose additionally on the hypothesis class; as the so-called sum to zero constraint, which makes both numerical implementation and theoretical analysis more involved.

To identify the *target function* to plug into the decoder, we fix a convex surrogate loss

$$\ell : \mathbb{R}^{T-1} \times \mathcal{Y} \to [0, \infty)$$

with corresponding risk $\mathcal{R}_\ell(f) = \mathbb{E}\ell(f(X), Y)$, and define

$$f_\ell = \arg\min_{f \in L^0(\mathcal{X}, \mathbb{R}^{T-1})} \mathcal{R}_\ell(f).$$

We then approximate f_ℓ by a (uniform) approximator f_λ. At the current level of generality, λ simply denotes a generic parameter to be tuned. For instance, f_λ can be the minimizer of a regularized risk, with λ the regularization parameter. Finally, our classifier will be $D\widehat{f}_\lambda$, with $\widehat{f}_\lambda$ the empirical estimate of f_λ based on the samples $\{(X_i, Y_i)\}_{i=1}^n$.

We are going to consider two cases of loss functions. The first case is the square loss $\ell(w, y) = \|w - y\|^2$, for which $f_\ell(x) = \eta(x)$, where

$$\eta(x) = \mathbb{E}[Y \mid X = x]$$

is the *regression function*. The second case is a family of functions of the *margin* $\langle w, y\rangle$, namely losses of the form $\ell_\phi(w, y) = \phi(\langle w, y\rangle)$ for a suitable (differentiable, convex) function $\phi : \mathbb{R} \to [0, \infty)$. Examples of ϕ are $\phi(t) = \ln(2)^{-1} \ln(1 + e^{-t})$ and $\phi(t) = e^{-t}$, generalizing the logistic and exponential loss to the multiclass setting, respectively. For margin losses, we will denote the minimizer f_ℓ by f_ϕ. Note that in binary classification the square loss is itself a function of the margin, $\ell(w, y) = \|1 - wy\|^2$, while for $T \geq 3$ this is no longer the case.

3. Analysis

We start by analyzing the peculiar structure of the bias-variance decomposition in classification. Note that such a decomposition is quite different from more classical decompositions such as (Geman et al., 1992). As we will see, the key point is that the bias can be made zero under suitable margin conditions. When only the variance is left, the misclassification error can be controlled by uniform concentration. These general facts can then be applied to different loss functions, leading to our main results.

3.1. Bias-variance for plug-in classifiers

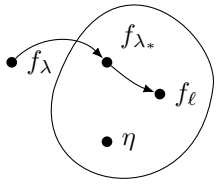
To analyze the performance of a plug-in classifier $D\widehat{f}_\lambda$, we decompose the excess misclassification risk as

$$\begin{aligned}\mathcal{R}(D\widehat{f}_\lambda) - \mathcal{R}_* &= \mathcal{R}(D\widehat{f}_\lambda) - \mathcal{R}(Df_\lambda) && (1)\\ &+ \mathcal{R}(Df_\lambda) - \mathcal{R}(Df_\ell) && (2)\\ &+ \mathcal{R}(Df_\ell) - \mathcal{R}_*. && (3)\end{aligned}$$

The last term results from replacing the 0-1 loss with the surrogate loss ℓ. Loss functions for which $Df_\ell = c_*$, and therefore (3) is zero, are called *Fisher consistent* (or *classification calibrated*). Fisher consistency is a common and well characterized property (Zhang, 2004a). In particular, the square loss is Fisher consistent (see Lemma 2). For margin losses, consistency will be assumed in all that follows, and shown in some examples.

The term (2) is a bias term. Crucially, it can be set to zero for a wide range of parameters λ. The idea is that we can have $\mathcal{R}(Df_\lambda) = \mathcal{R}(Df_\ell)$ even when $\mathcal{R}_\ell(f_\lambda) \gg \mathcal{R}_\ell(f_\ell)$. Here is a fundamental difference between regression and classification. While in regression f_ℓ is a target point, in classification it is rather a representative of the target class

$$[f_\ell] = \{f \in L^0(\mathcal{X}, \mathbb{R}^{T-1}) : Df = Df_\ell \text{ almost surely}\}.$$



Hence, it is enough for f_λ to land in $[f_\ell]$, possibly far from f_ℓ itself. This is easier if the class $[f_\ell]$ is "large", which can be ensured by imposing special margin conditions. Assuming that ℓ is Fisher consistent, a generic function f lies in $[f_\ell]$ if and only if

$$Df = c_* \quad \text{almost surely.} \tag{4}$$

Chosen a Fisher consistent loss and put the bias to zero, all that's left is the variance term (1):

$$\mathcal{R}(D\widehat{f}_\lambda) - \mathcal{R}_* = \mathcal{R}(D\widehat{f}_\lambda) - \mathcal{R}(Df_\lambda). \tag{5}$$

At this point, λ is set and needs no trade-off. Fast convergence of the variance, and therefore of the whole excess misclassification risk, can be derived using once again margin conditions.

3.2. Margin conditions

In binary classification, the *margin conditions*, also known as Tsybakov's low-noise assumptions (Mammen & Tsybakov, 1999; Tsybakov, 2004; Koltchinskii & Beznosova, 2005; Audibert & Tsybakov, 2007), are a set of assumptions under which it is possible to obtain fast convergence (up to exponential) for plug-in classifiers. They can be stated as follows: there exists $\alpha \in (0, \infty]$ such that, for every $\delta > 0$,

$$\mathbb{P}\{|\eta(X)| \leq \delta\} \lesssim \delta^\alpha. \tag{6}$$

In the extreme case of $\alpha = \infty$, we get

$$|\eta(X)| \geq \delta \quad \text{almost surely,} \tag{7}$$

which is sometimes referred to as the *hard-margin* condition.

Following (Mroueh et al., 2012; Nowak et al., 2019), we can generalize (6) and (7) to the multiclass setting. For $w \in \mathbb{R}^{T-1}$, we define the *decision margin*

$$M(w) = \min_{y \neq D(w)} \langle w, D(w) - y\rangle.$$

$M(w)$ is the difference between the largest and the second largest projection of w onto $\mathcal{Y}$, namely the confidence gap between first and second guess. For $T = 2$, we have $M(w) = 2|w|$. In general, we say that a function $f : \mathcal{X} \to \mathbb{R}^{T-1}$ satisfies the margin condition with exponent $\alpha \in (0, \infty]$ if, for every $\delta > 0$,

$$\mathbb{P}\{M(f(X)) \leq \delta\} \lesssim \delta^\alpha. \tag{8}$$

In particular, one can take $f = \eta$, which for $T = 2$ gives back (6). Again, $\alpha = \infty$ gives the hard-margin condition (see Figure 3)

$$M(f(X)) \geq \delta \quad \text{almost surely.} \tag{9}$$

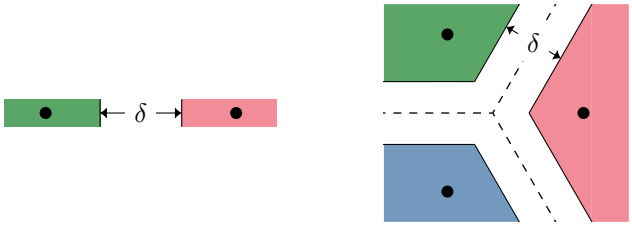


Figure 3. Hard-margin condition ($T = 2, 3$).

Intuitively, these conditions say that the probability of falling in a "runoff zone", where the plugging-in function would be "uncertain", is either (polynomially) small (8), or zero (9). The reason why we state (8) and (9) for an arbitrary f is that we will transfer these properties to minimizers f_ℓ (and f_λ) of general (regularized) losses, including but not limited to the square. Combining Fisher consistency and hard margin, we obtain the following stronger condition.

Lemma 1. *A function $f \in L^0(\mathcal{X}, \mathbb{R}^{T-1})$ satisfies (4) and (9) if and only if*

$$\min_{y \neq c_*(X)} \langle f(X), c_*(X) - y \rangle \geq \delta \quad \text{almost surely.} \tag{10}$$

Proof. First note that, if (4) holds, then (10) is the same as (9). Now suppose (10) holds. Then

$$\langle f(X), c_*(X) \rangle > \max_{y \neq c_*(X)} \langle f(X), y \rangle,$$

hence $Df(X) = c_*(X)$, that is, (4) holds too. □

Beside (8) and (9), we will also consider another generalization of the hard margin (7) which is independent of any particular classifier, and instead is stated purely in terms of the conditional probabilities. To illustrate such a condition, we note that (7) is equivalent to saying that either $\rho(1|x)$ or $\rho(-1|x)$ is no less than $1/2 + \delta/2$ (for almost every x, there is one class with probability bounded away from coin flipping). This in turn is equivalent to

$$\min_{y \neq c_*(X)} \rho(c_*(X) \mid X) - \rho(y \mid X) \geq \delta \quad \text{almost surely,} \tag{11}$$

which says that the most probable class has almost always an edge of δ over the second most probable class. Since this inequality makes sense for arbitrary T, we take it as our hard-margin condition for multiclass problems. More generally, one may consider problems where for some $\alpha \in (0, \infty]$ and all $\delta > 0$,

$$\mathbb{P}\{\min_{y \neq c_*(X)} \rho(c_*(X) \mid X) - \rho(y \mid X) \leq \delta\} \lesssim \delta^\alpha, \tag{12}$$

generalizing (6) to $T \geq 2$.

The following Lemma describes the relationship between margin on conditional probabilities and margin on classifiers.

Lemma 2. *We have $D\eta = c_*$ almost surely. Moreover, (11) holds if and only if η satisfies (9).*

Proof. Let

$$\Delta = \{p \in \mathbb{R}^{\mathcal{Y}} : p_y \geq 0, \sum_{y \in \mathcal{Y}} p_y = 1\}$$

be the probability simplex on $\mathcal{Y}$, and let $\operatorname{co} \mathcal{Y}$ be the encoding simplex defined as the convex hull of $\mathcal{Y}$. Then Δ and $\operatorname{co} \mathcal{Y}$ are canonically isomorphic via the barycenter coordinate map

$$\beta : \Delta \to \operatorname{co} \mathcal{Y}, \qquad \beta(p) = \sum_{y \in \mathcal{Y}} p_y y.$$

Now consider the map

$$\rho : \mathcal{X} \to \Delta, \qquad \rho(x) = [\rho(y \mid x)]_{y \in \mathcal{Y}}.$$

Then we have $\beta \circ \rho = \eta$. It follows that $y \in \mathcal{Y}$ maximizes $\rho(y \mid x)$ if and only if it maximizes $\langle \eta(x), y \rangle$. Therefore, $D\eta = c_*$. The same holds for maximizing over $y \neq c_*(x)$, whence the second claim. □

3.3. Misclassification comparison

In view of (5), we need in fact to compare the misclassification risk of two classifiers. This can be done by introducing a bounding distance. Since the distance will be symmetric, the resulting bound will give a symmetric comparison between any two classifiers, as opposed to the usual comparison of a classifier with respect to a fixed (Bayes) rule. For this reason, the following results may be of independent interest.

We define the *Hamming distance* of $c', c \in L^0(\mathcal{X}, \mathcal{Y})$ as

$$r(c', c) = \mathbb{P}\{c'(X) \neq c(X)\}.$$

The Hamming distance bounds the difference of misclassification risk.

Lemma 3. *For every $c', c \in L^0(\mathcal{X}, \mathcal{Y})$,*

$$|\mathcal{R}(c') - \mathcal{R}(c)| \leq r(c', c).$$

Proof. By direct computation,

$$\begin{aligned}|\mathcal{R}(c') - \mathcal{R}(c)| &= |\mathbb{E}[\mathbb{1}\{c'(X) \neq Y\} - \mathbb{1}\{c(X) \neq Y\}]| \\ &\leq \mathbb{E}[|\mathbb{1}\{c'(X) \neq Y\} - \mathbb{1}\{c(X) \neq Y\}|] \\ &\leq \mathbb{E}[\mathbb{1}\{c'(X) \neq c(X)\}] = r(c', c). \qquad \square\end{aligned}$$

The next step is to bound the Hamming distance between two plug-in classifiers.

Lemma 4. *For every $f', f \in L^\infty(\mathcal{X}, \mathbb{R}^{T-1})$,*

$$r(Df', Df) \leq \mathbb{P}\left\{\|f' - f\|_\infty \geq \sqrt{\tfrac{T-1}{2T}} M(f(X))\right\}.$$

Proof. Let $Df'(x) = y' \neq y = Df(x)$. Then

$$\begin{aligned}\min_{j \neq y'} \langle y' - j, f'(x)\rangle &= \langle y', f'(x)\rangle - \max_{j \neq y'} \langle j, f'(x)\rangle \\ &\leq \langle y', f'(x)\rangle - \langle y, f'(x)\rangle \\ &\leq \langle y' - y, f'(x)\rangle - \langle y' - y, f(x)\rangle \\ &= \langle y' - y, f'(x) - f(x)\rangle \\ &\leq \|y' - y\| \|f'(x) - f(x)\| \\ &\leq \sqrt{\tfrac{2T}{T-1}} \|f' - f\|_\infty. \qquad \square\end{aligned}$$

Now we let the samples come into play. Let $\widehat{f} \in L^\infty(\mathcal{X}, \mathbb{R}^{T-1})$ be a function of (X_i, Y_i), $i = 1, \ldots, n$, such that, for every $\epsilon > 0$ and some constant $b > 0$,

$$\mathbb{P}\{\|\widehat{f} - f\|_\infty > \epsilon\} \lesssim \exp(-n\epsilon^2/b^2). \tag{13}$$

Then the following polynomial and exponential bounds hold true.

Proposition 5. *Suppose f satisfies the margin condition (8), and let $\widehat{f}$ obey the concentration (13). Then*

$$\mathbb{E}|\mathcal{R}(D\widehat{f}) - \mathcal{R}(Df)| \lesssim b^\alpha \left(\tfrac{2T}{T-1}\right)^{\alpha/2} \left(\tfrac{\log n^{\alpha/2}}{n}\right)^{\alpha/2}.$$

If f satisfies the hard-margin condition (9), then

$$\mathbb{E}|\mathcal{R}(D\widehat{f}) - \mathcal{R}(Df)| \lesssim \exp(-n\delta^2/b^2).$$

Proof. By Lemma 3 and Lemma 4,

$$\begin{aligned}&\mathbb{E}|\mathcal{R}(D\widehat{f}) - \mathcal{R}(Df)| \leq \mathbb{E} r(D\widehat{f}, Df) \\ &\leq \mathbb{E}_{\{(X_i,Y_i)\}_{i=1}^n} \mathbb{E}_X \mathbb{1}\left\{\|\widehat{f} - f\|_\infty \geq \sqrt{\tfrac{T-1}{2T}} M(f(X))\right\}.\end{aligned}$$

Let $\gamma = \sqrt{\frac{T-1}{2T}} M(f(X))$ and $E = \{M(f(X)) \leq \delta\}$. Then we have

$$\begin{aligned}&\mathbb{E}_X \mathbb{1}\{\|\widehat{f} - f\|_\infty \geq \gamma\} \\ &= \mathbb{E}_X[\mathbb{1}\{\|\widehat{f} - f\|_\infty \geq \gamma\} \mid E\,] \,\mathbb{P}\{E\} \\ &+ \mathbb{E}_X[\mathbb{1}\{\|\widehat{f} - f\|_\infty \geq \gamma\} \mid E^{\complement}\,] \,\mathbb{P}\{E^{\complement}\} \\ &\leq \mathbb{P}\{E\} + \mathbb{1}\{\|\widehat{f} - f\|_\infty \geq \sqrt{\tfrac{T-1}{2T}}\delta\},\end{aligned}$$

where $\mathbb{P}\{E\} \lesssim \delta^\alpha$ by (8). Moreover, thanks to (13),

$$\begin{aligned}&\mathbb{E}_{\{(X_i,Y_i)\}_{i=1}^n} \mathbb{1}\{\|\widehat{f} - f\|_\infty \geq \sqrt{\tfrac{T-1}{2T}}\delta\} \\ &= \mathbb{P}\left\{\|\widehat{f} - f\|_\infty \geq \sqrt{\tfrac{T-1}{2T}}\delta\right\} \\ &\lesssim \exp(-n\tfrac{T-1}{2T}\delta^2/b^2).\end{aligned}$$

Setting $\delta^2 = b^2 \frac{2T}{T-1}(\log(n^{\alpha/2})/n)$, we obtain the first claimed inequality. The second inequality follows similarly using (9) in place of (8). $\square$

3.4. Main results

In this section we establish exponential convergence of plug-in classifiers under assumptions of hard margin. We assume the setting of Section 2, and use the arguments of Sections 3.1 to 3.3. The main results are given for two cases of loss functions, first for the square loss (namely, for the regression function), and then for a general family of margin losses. We will also be making the additional assumptions below.

(i) $f_\ell \in L^\infty(\mathcal{X}, \mathbb{R}^{T-1})$ and $\|f_\lambda - f_\ell\|_\infty \xrightarrow[\lambda \to 0]{} 0$.

The limit in (i) is taken with respect to the amount of (explicit or implicit) regularization. For example, λ can represent a Tikhonov parameter, the reciprocal of the number of iterations in gradient descent, or the reciprocal of the number of weights in a neural network model.

Further, let $\widehat{f}_\lambda \in L^\infty(\mathcal{X}, \mathbb{R}^{T-1})$ be an estimate of f_λ, and assume that, for every λ and some $b_\lambda > 0$, the following concentration bound holds true:

(ii) $\mathbb{P}\{\|\widehat{f}_\lambda - f_\lambda\|_\infty > \epsilon\} \lesssim \exp(-n\epsilon^2/b_\lambda^2)$.

Regularization methods in reproducing kernel Hilbert spaces (RKHS) (Schölkopf & Smola, 2002; Steinwart & Christmann, 2008) provide one framework where the properties (i), (ii) can be satisfied. In particular, one can fix a separable RKHS $\mathcal{H} \subset L^0(\mathcal{X}, \mathbb{R}^{T-1})$ with norm $\|\cdot\|_\mathcal{H}$, and define

$$f_\lambda = \operatorname*{arg\,min}_{f \in \mathcal{H}} \mathcal{R}_\ell(f) + \lambda \|f\|_\mathcal{H}^2, \quad \lambda \geq 0.$$

If $\mathcal{H}$ has kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that $\sup_x K(x, x) \leq \kappa^2$, $\mathcal{H}$ is continuously embedded in the space of bounded continuous functions on $\mathcal{X}$, with $\|\cdot\|_\infty \leq \kappa \|\cdot\|_\mathcal{H}$. Hence, the uniform bounds (i), (ii) may be derived from bounds in the RKHS norm. The estimate $\widehat{f}_\lambda$ can be computed with a variety of methods, such as empirical risk minimization (ERM) (Schölkopf & Smola, 2002), gradient descent (GD) (Yao et al., 2007) and stochastic gradient descent (SGD) (Robbins & Monro, 1951).

Lemma 6. *Suppose (9) holds true with $f = f_\ell$ and $\delta = \gamma$ for some $\gamma > 0$. Then, under the assumption (i), there is λ_* such that (10) holds true with $f = f_\lambda$ and $\delta = \gamma/2$ for every $\lambda \leq \lambda_*$.*

Proof. Let $Df_\ell(x) = y_* = c_*(x)$ (recall that ℓ is Fisher consistent), and let

$$a = \langle f_\lambda(x), y_* \rangle - \max_{y \neq y_*} \langle f_\lambda(x), y \rangle.$$

Then

$$a = \langle f_\ell(x), y_* \rangle - \underbrace{\langle f_\ell(x) - f_\lambda(x), y_* \rangle}_{b} - \underbrace{\max_{y \neq y_*} \langle f_\lambda(x), y \rangle}_{c},$$

where $b \leq \|f_\ell - f_\lambda\|_\infty$, and

$$\begin{aligned} c &= \max_{y \neq y_*} (\langle f_\ell(x), y \rangle + \langle f_\lambda(x) - f_\ell(x), y \rangle) \\ &\leq \max_{y \neq y_*} \langle f_\ell(x), y \rangle + \max_{y \neq y_*} \langle f_\lambda(x) - f_\ell(x), y \rangle \\ &\leq \max_{y \neq y_*} \langle f_\ell(x), y \rangle + \|f_\lambda - f_\ell\|_\infty. \end{aligned}$$

In view of (i), there is λ_* such that $\|f_\lambda - f_\ell\|_\infty \leq \gamma/4$. Hence, for every $\lambda \leq \lambda_*$, using (9) we obtain

$$\begin{aligned} a &\geq \langle f_\ell(x), y_* \rangle - \gamma/4 - \max_{y \neq y_*} \langle f_\ell(x), y \rangle - \gamma/4 \\ &= M(f_\ell(x)) - \gamma/2 \geq \gamma - \gamma/2 = \gamma/2. \end{aligned}$$

This implies (4), and thus (9), for $f = f_\lambda$ and $\delta = \gamma/2$. The assertion now follows from Lemma 1. □

Square loss. We can now state our first main result.

Theorem 7. *Suppose the hard-margin condition*

$$\min_{y \neq c_*(X)} \rho(c_*(X) \mid X) - \rho(y \mid X) \geq \delta \quad \textit{almost surely}.$$

Then, under the assumptions (i) and (ii), there is λ_* such that, for every $\lambda \leq \lambda_*$,

$$\mathbb{E}|\mathcal{R}(D\widehat{f}) - \mathcal{R}_*| \lesssim \exp(-n\delta^2\lambda/b_\lambda^2).$$

Proof. First, recall that, thanks to Lemma 2, $D\eta = c_*$ and $M(\eta(X)) \geq \delta$ almost surely. Moreover, by Lemma 6 (and Lemma 1), we have $Df_\lambda = c_*$ and $M(f_\lambda(X)) \geq \delta/2$ almost surely for $\lambda \leq \lambda_*$. Thus, (5) holds true, and the claim follows from Proposition 5. □

Margin losses. We now consider surrogate losses of the form

$$\ell_\phi(w, y) = \phi(\langle w, y \rangle)$$

for some scalar function $\phi : \mathbb{R} \to [0, \infty)$. We denote the minimizer of the corresponding risk $\mathcal{R}_\phi(f) = \mathbb{E}\phi(\langle f(X), Y \rangle)$ by

$$f_\phi = \operatorname*{arg\,min}_{f \in L^0(\mathcal{X}, \mathbb{R}^{T-1})} \mathcal{R}_\phi(f).$$

Following and generalizing the analysis of (Zhang, 2004a; Nitanda & Suzuki, 2019), we want to extract an inner risk from $\mathcal{R}_\phi$. The idea is to expand

$$\mathcal{R}_\phi(f) = \mathbb{E}_X \sum_{y \in \mathcal{Y}} \phi(\langle f(X), y \rangle) \rho(y \mid X)$$

and isolate the argument of $\mathbb{E}_X$ removing the dependence on X. Recalling the definition of Δ in Lemma 2, we introduce the inner risk

$$\Phi(p, w) = \sum_{y \in \mathcal{Y}} \phi(\langle w, y \rangle) p_y, \qquad p \in \Delta, w \in \mathbb{R}^{T-1},$$

and the inner risk minimizer

$$h_\phi : \Delta \to \mathbb{R}^{T-1}, \qquad h_\phi(p) = \operatorname*{arg\,min}_{w \in \mathbb{R}^{T-1}} \Phi(p, w).$$

Note that, denoting $p(x)_y = \rho(y \mid x)$, we have

$$f_\phi(x) = h_\phi(p(x)). \tag{14}$$

In the following, we will be assuming that

(iii) ℓ_ϕ is Fisher consistent;

(iv) $\langle h_\phi(p), y \rangle$ is a non-decreasing function of p_y .

As previously mentioned, losses satisfying (iii) are indeed abundant. For a general characterization of Fisher consistency in the framework of simplex encoded classification, we refer to (Mroueh et al., 2012). The requirement (iv) is easily met by many functions ϕ, as the next lemma shows. Essentially, it is sufficient for the loss to be decreasing and convex. Notable examples of ϕ satisfying both (iii) and (iv) are the logistic loss $\phi(t) = \ln(2)^{-1} \ln(1 + e^{-t})$, and the exponential loss $\phi(t) = e^{-t}$.

Lemma 8. *Suppose ϕ is twice differentiable, non-increasing and convex. Then (iv) holds true.*

Proof. Let $\Psi(p) = \nabla_w \Phi(p, h_\phi(p))$ be the gradient with respect to Φ's second argument. Computing the Jacobian of $\Psi(p)$ with respect to p we have

$$\begin{aligned} \mathrm{J}\Psi(p) = \sum_{j \in \mathcal{Y}} \Big(&\phi''(\langle h_\phi(p), j \rangle)\, j \otimes j\ \mathrm{J}h_\phi(p) p_j \\ &+ \phi'(\langle h_\phi(p), j \rangle)\, j \otimes e_j \Big), \end{aligned}$$

where e_j denotes the vector of $\mathbb{R}^T$ with $[e_j]_y = \delta_{j,y}$. By definition of $h_\phi(p)$, both $\Psi(p) = 0$ and $\mathrm{J}\Psi(p) = 0$. Thus, for all $y \in \mathcal{Y}$,

$$0 = \mathrm{J}\Psi(p)e_y = \sum_{j\in\mathcal{Y}} p_j\ \phi''(\langle h_\phi(p), j\rangle)\ j\ \langle \tfrac{\partial h_\phi(p)}{\partial p_y}, j\rangle + \phi'(\langle h_\phi(p), y\rangle)\ y,$$

and therefore

$$\begin{aligned} 0 &= \langle \tfrac{\partial h_\phi(p)}{\partial p_y}, \mathrm{J}\Psi(p)e_y\rangle \\ &= \sum_{j\in\mathcal{Y}} p_j\ \phi''(\langle h_\phi(p), j\rangle)\ \langle \tfrac{\partial h_\phi(p)}{\partial p_y}, j\rangle^2 \\ &\quad + \phi'(\langle h_\phi(p), y\rangle)\ \langle \tfrac{\partial h_\phi(p)}{\partial p_y}, y\rangle. \end{aligned}$$

Since $\phi'' \geq 0$ and $\phi' \leq 0$, we must have $\langle \frac{\partial h_\phi(p)}{\partial p_y}, y\rangle \geq 0$, which proves the claim. □

In order to derive exponential rates for margin losses, we need to transfer the hard-margin condition from the conditional probabilities to the minimizer of the margin loss. This is the content of the following lemma.

Lemma 9. *Suppose that* (11) *holds true with* $\delta = \gamma$. *Then, under the assumption* (iv), (9) *holds true with* $f = f_\phi$ *and* $\delta = m(\gamma)$, *where*

$$m(\gamma) = \max_{y,j\in\mathcal{Y}} \min\{M(h_\phi(p)) : p \in \Delta, p_y - p_j = 2\gamma\}.$$

Proof. Let $p(X)_y = \rho(y \mid X)$. By (14) we have

$$M(f_\phi(X)) = M(h_\phi(p(X))).$$

Let $y, j \in \mathcal{Y}$ be such that

$$M(h_\phi(p(X))) = \underbrace{\langle h_\phi(p(X)), y\rangle}_{a} - \underbrace{\langle h_\phi(p(X)), j\rangle}_{b}.$$

In view of (11) and (iv), there is $p \in \Delta$ with $p_y - p_j = 2\delta$ such that a decreases and b increases, hence

$$M(h_\phi(p(X))) \geq M(h_\phi(p)).$$

Taking the minimum over such a p and the maximum over y and j, we obtain the assertion. □

To visualize the lower bound $m(\gamma)$, note that, for $T = 2$, it corresponds to $\max\{h_\phi(1/2+\gamma), -h_\phi(1/2-\gamma)\}$ (cf. with Nitanda & Suzuki (2019)).

We can finally prove our main result for margin losses.

Theorem 10. *Suppose the hard-margin condition*

$$\min_{y\neq c_*(X)} \rho(c_*(X) \mid X) - \rho(y \mid X) \geq \delta \quad \textit{almost surely}.$$

Then, under the assumptions (i), (ii), (iii) *and* (iv), *there is* λ_* *such that, for every* $\lambda \leq \lambda_*$,

$$\mathbb{E}|\mathcal{R}(D\widehat{f}) - \mathcal{R}_*| \lesssim \exp(-n\ m(\delta)^2\lambda/b_\lambda^2),$$

where $m(\delta)$ *is defined in Lemma 9.*

Proof. By assumption (iii), we have $Df_\phi = c_*$ almost surely. Moreover, thanks to Lemma 9, we have $M(f_\phi(X)) \geq m(\delta)$ almost surely. Now, Lemma 6 (together with Lemma 1) gives that $Df_\lambda = c_*$ and $M(f_\lambda(X)) \geq m(\delta)/2$ almost surely for $\lambda \leq \lambda_*$. Therefore, we have (5), and Proposition 5 yields the result. □

The critical value λ_* in Theorem 7 and Theorem 10 can be quantified in presence of additional assumptions on the distribution. For example, consider the case of a kernel ridge regression estimator in a separable RKHS $\mathcal{H}$. Suppose that the kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is bounded by κ, and define the covariance operator as

$$T : \mathcal{H} \to \mathcal{H}, \qquad T = \int_{\mathcal{X}} K_x \otimes K_x d\rho(x),$$

where $K_x = K(\cdot, x)$ and ρ is the marginal distribution on $\mathcal{X}$. Further, suppose there exist $g \in \mathcal{H}$ and $s \in (0, 1/2]$ such that

$$f_\ell = T^s g.$$

This is known as the *source* condition, and it corresponds to assuming Sobolev smoothness of the regression function. Then, it can be proved that (Caponnetto & De Vito, 2007)

$$\|f_\lambda - f_\ell\|_{\mathcal{H}} \leq \lambda^s \|g\|_{\mathcal{H}}.$$

As a consequence, λ_* in Lemma 6, and therefore in Theorem 7, may be picked as $\lambda_* = (\delta/4\kappa\|g\|_{\mathcal{H}})^{1/s}$.

We finally remark that analogous results to Theorem 7 and Theorem 10 may be proved under the soft-margin condition (12), using the polynomial bound of Proposition 5.

4. Experiments

This section is concerned with empirically verifying the theoretical analysis presented in Section 3. We will first consider a classification problem in which the true function satisfies the hard-margin condition (defined in (9)), and show how – under optimization of a surrogate loss by gradient descent – the misclassification loss decreases faster than the surrogate loss. Then we will take into account a different synthetic dataset, where the weaker soft-margin or low noise condition (see (8)) is satisfied. We will verify how the rate of change of the misclassification error with the number of data-points adheres to the theoretical rates.

Initially we compare the logistic, exponantial and square surrogate loss functions. A synthetic, two-dimensional dataset

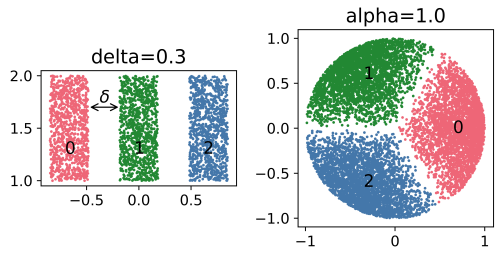


Figure 4. Sample datasets. In the left panel, the three classes are separated by a hard margin of length δ. In the right panel there is no hard margin, but the probability of a point falling close to the boundary is decreasing (soft-margin).

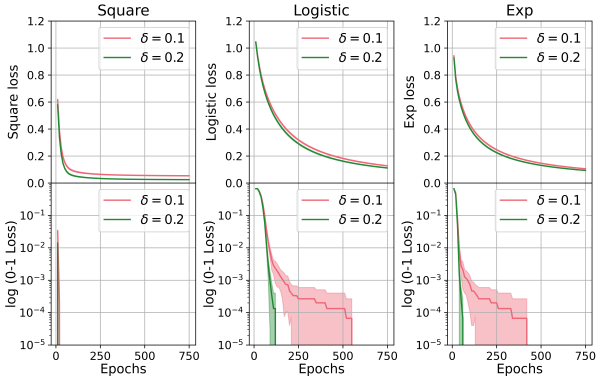


Figure 5. Optimization curves on hard-margin classification with different surrogate losses. Each panel contains two curves calculated on datasets with different margins δ. The top row shows the surrogate loss, the bottom row shows the 0-1 loss.

is generated such that the hard-margin condition holds with margin δ, see Figure 4, left panel for a sample dataset. Random Fourier features (RFF) models (Rahimi & Recht, 2008) approximate potentially infinite dimensional feature maps in a reproducing kernel Hilbert space (RKHS) using finite dimensional randomized maps: given a kernel function $k(x, x') = \langle \psi(x), \psi(x') \rangle_{\mathcal{H}}$ the feature map $\psi \in \mathcal{H}$ can be approximated with function $z : \mathbb{R}^D \to \mathbb{R}^R$ such that $\langle \psi(x), \psi(x') \rangle_{\mathcal{H}} \approx \langle z(x), z(x') \rangle_{\mathbb{R}}$. Finally $z(x)$ can be used instead of the sample itself in a linear model with parameters $w \in \mathbb{R}^R$: $f(x) = w^\top z(x)$. In order to learn the parameters w we minimize the regularized surrogate loss with gradient descent. In Figure 5 we plot the 0-1 error and the surrogate losses on unseen data as a function of the optimization step. The experiments were repeated 20 times to obtain error bands. Note how the 0-1 loss converges at a faster pace than the surrogate. When the 0-1 loss is zero our classifier has entered the region in which its *decoded* value is equal to the true classifier, even if the surrogate loss is still far from zero. We can further notice how not all surrogates are equal: for both the small ($\delta = 0.1$) and the larger margin ($\delta = 0.2$), the square loss leads to faster convergence of the 0-1 error than both exponential and logistic losses.

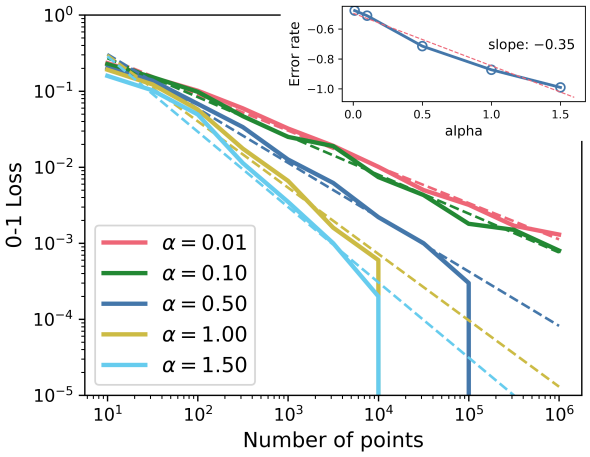


Figure 6. Main figure: error rates for multiclass classification with polynomial soft-margin with increasing dataset size. *Inset*: Linear rate of convergence of the error with α.

The second experiment was carried out with a 2D synthetic dataset comprising of three classes, such that the probability of a point falling close to the decision boundary decreases with the distance to the boundary itself as M^α for margins $M < 1$ (see (8)). A linear model was trained on this dataset by minimizing the regularized logistic loss with gradient descent until convergence. The experiment was repeated 100 times using different values of α (a higher α equates to an easier problem) and an increasing number of training points, recording the average 0-1 loss over unseen data. By plotting the 0-1 loss against the number of training samples for each α, we could observe an approximately linear trend on a log-log scale (see Figure 6). From Proposition 5 the error is expected to drop more rapidly with higher α, and in particular the rate of decrease should be $n^{-\alpha/2}$ (ignoring constant and logarithmic factors). By plotting the slopes of the error rates we obtain a straight line with slope -0.35, which is close to the prediction of -0.5 (see the inset on Figure 6).

5. Conclusions

In this paper we have shown how, under the hard-margin condition and for a very general framework which encompasses many different models and surrogate losses, the multiclass classification error exhibits exponentially fast convergence. Along the way we have provided an error decomposition where the bias term disappears. This kind of result fits with the recent empirical observations of how even highly overparametrized models do not overfit the training data. Our analysis can be experimentally verified for several losses, and different margin conditions.

Several possible extensions of this work have been left for future work. Beyond the hard-margin and low-noise conditions, robustness with respect to different kinds of noise may be studied. The explicit application of our bounds to specific models – which was sketched in this paper for kernel ridge regression – could be especially interesting for (deep) neural networks, for which fast convergence on classification problems has been ascertained. Indeed, for the latter models, the interplay of exponential convergence and overparameterization is a further topic of great interest.

Acknowledgements

L. R. acknowledges the financial support of the European Research Council (grant SLING 819789), the AFOSR project FA9550-18-1-7009 (European Office of Aerospace Research and Development), the EU H2020-MSCA-RISE project NoMADS - DLV-777826, and the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216.

References

Audibert, J.-Y. and Tsybakov, A. B. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608 – 633, 2007.

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

Cabannes, V. A., Bach, F., and Rudi, A. Fast rates for structured prediction. In *34th Conference on Learning Theory*, volume 134, pp. 823–865, 2021.

Caponnetto, A. and De Vito, E. Optimal Rates for the Regularized Least-Squares Algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Chen, D.-R. and Sun, T. Consistency of Multiclass Empirical Risk Minimization Methods Based on Convex Loss. *Journal of Machine Learning Research*, 7(86):2435–2447, 2006.

Feldman, V., Guruswami, V., Raghavendra, P., and Wu, Y. Agnostic learning of monomials by halfspaces is hard. *SIAM Journal on Computing*, 41(6):1558–1590, 2012.

Geman, S., Bienenstock, E., and Doursat, R. Neural Networks and the Bias/Variance Dilemma. *Neural Computation*, 4(1):1–58, 1992.

Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, 2009.

Koltchinskii, V. and Beznosova, O. Exponential Convergence Rates in Classification. In *International Conference on Computational Learning Theory*, pp. 295–307, 2005.

Mammen, E. and Tsybakov, A. B. Smooth Discrimination Analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999.

Mroueh, Y., Poggio, T., Rosasco, L., and Slotine, J.-j. Multiclass Learning with Simplex Coding. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

Nitanda, A. and Suzuki, T. Stochastic Gradient Descent with Exponential Convergence Rates of Expected Classification Errors. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1417–1426. PMLR, 2019.

Nowak, A., Bach, F., and Rudi, A. Sharp analysis of learning with discrete losses. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1920–1929. PMLR, 2019.

Pillaud-Vivien, L., Rudi, A., and Bach, F. Exponential convergence of testing error for stochastic gradient methods. In *The 31st Conference On Learning Theory*, pp. 250–296, 2018.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, 2008.

Robbins, H. and Monro, S. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3): 400 – 407, 1951.

Schölkopf, B. and Smola, A. J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.

Steinwart, I. and Christmann, A. *Support Vector Machines*. Information Science and Statistics. Springer New York, 2008.

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

Wolpert, D. H. The Lack of A Priori Distinctions Between Learning Algorithms. *Neural Computation*, 8(7):1341–1390, 1996.

Yao, Y., Rosasco, L., and Caponnetto, A. On early stopping in gradient descent learning. *Constructive Approximation*, 26:289–315, 2007.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Zhang, T. Statistical Behavior and Consistency of Classification Methods Based on Convex Risk Minimization. *The Annals of Statistics*, 32(1):56–85, 2004a.

Zhang, T. Statistical Analysis of Some Multi-Category Large Margin Classification Methods. *Journal of Machine Learning Research*, 5:1225–1251, 2004b.