Computational Complexity of Linear Large Margin Classification With Ramp Loss

Søren Frejstrup Maibing
Department of Mathematical Sciences
University of Copenhagen

Christian Igel
Department of Computer Science
University of Copenhagen

Abstract

Minimizing the binary classification error with a linear model leads to an NP-hard problem. In practice, surrogate loss functions are used, in particular loss functions leading to large margin classification such as the hinge loss and the ramp loss. The intuitive large margin concept is theoretically supported by generalization bounds linking the expected classification error to the empirical margin error and the complexity of the considered hypotheses class. This article addresses the fundamental question about the computational complexity of determining whether there is a hypotheses class with a hypothesis such that the upper bound on the generalization error is below a certain value. Results of this type are important for model comparison and selection. This paper takes a first step and proves that minimizing a basic margin-bound is NP-hard when considering linear hypotheses and the ρ-margin loss function, which generalizes the ramp loss. This result directly implies the hardness of ramp loss minimization.

1 INTRODUCTION

Some of the most fundamental problems in machine learning are NP-hard (Johnson and Preparata, 1978; Hush, 1999; Šíma, 2002; Bartlett and Ben-David, 2002; Ben-David et al., 2003b; Feldman et al., 2009). Let us consider binary classification based on sample data $S = \{(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_\ell, y_\ell)\}$ drawn i.i.d. from an unknown joint distribution D over the input space $\mathcal{X} = \mathbb{R}^n$ and the output space $\mathcal{Y} = \{-1, 1\}$. Our goal is to find a hypothesis $h : \mathcal{X} \to \mathbb{R}$ minimizing the generalization error (risk) $R^{0\text{-}1}(h) = \mathbb{E}_{(\boldsymbol{x},y)\sim D}[L_{0\text{-}1}(y, h(\boldsymbol{x}))]$ under the 0-1 loss $L_{0\text{-}1}(y, y') = \mathbb{1}_{y \neq \operatorname{sgn}(y')}$, where the indicator function $\mathbb{1}_A$ returns 1 if A is true and 0 otherwise and $\operatorname{sgn}(z) = 2\mathbb{1}_{z\geq 0} - 1$. The generalization error can be estimated by the empirical risk $R_S^{0\text{-}1}(h) := \frac{1}{\ell}\sum_{i=1}^{\ell} L_{0\text{-}1}(y_i, h(\boldsymbol{x}_i))$. However, even if we restrict h to be of affine linear form, minimizing the empirical risk is NP hard (the size of the problem is measured by the number of bits describing S, w.l.o.g. assuming rational input values). Thus, one typically replaces the 0-1 loss by a surrogate loss function L_{surr} (e.g., the squared loss, the ramp loss, or the hinge loss). Then generalization bounds are derived linking the generalization error to the empirical risk under the surrogate loss, the number of data points, and restrictions on h (Boucheron et al., 2005). These bounds are usually of the form

$$R^{0\text{-}1}(h) \leq R_S^{\text{surr}}(h) + B(H, \ell, \delta)$$

with probability $1-\delta$, where H denotes a restricted hypotheses class, $h \in H$, and the term B is decreasing in ℓ and δ. Such bounds do not only provide a better theoretical understanding, they also motivate the design of new learning algorithms. They are used (and sometimes abused) for comparing hypotheses and as a basis for model selection strategies.

Given that $R_S^{\text{surr}}(h)$, $h \in H$, can be minimized efficiently, it seems that the computational hardness problem has been circumvented. However, ultimately we strive for hypotheses with small

Appearing in Proceedings of the 18th International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38.

risk. The expected risk is assessed by $R_S^{\mathrm{surr}}(h) + B(H, \ell, \delta)$, where both terms depend on the choice of the hypotheses class H. Tuning H to the problem at hand is an optimization of its own and often referred to as model selection. For soft-margin support vector machines (SVMs, Cortes and Vapnik, 1995; Schölkopf and Smola, 2002), the most prominent large margin classifiers, this amounts to choosing a regularization parameter and, in the non-linear case, a kernel function. Now, the—in our opinion fundamental and so far mostly neglected—question arises: what is the computational complexity of determining if $R_S^{\mathrm{surr}}(h) + B(H, \ell, \delta)$ can be smaller than a certain value for some choice of H? Results of this type are necessary for the analysis of model comparison and selection techniques relying on generalization bounds. In this study, we take a first step and consider the computational complexity of minimizing a margin bound for linear hypotheses.

Large margin classification, which aims at maximizing the minimum distance of correctly classified training input points from the decision boundary between the classes, is a common strategy for finding hypotheses that generalizes well. It can be motivated by minimizing the right hand side of generalization bounds (Vapnik, 1998; Schölkopf and Smola, 2002; Boucheron et al., 2005; Mohri et al., 2012). Typical loss functions leading to large margin classification are the *hinge loss* and the *ρ-margin loss function* (Mohri et al., 2012). In basic generalization bounds for large margin classifiers, often called margin bounds, B depends on the margin in relation to the spread of the training data as measured by, for example, the trace of the Gram matrix or the radius of the smallest ball containing the training data.

In the following, we focus on the (non-convex) ρ-margin loss function (Mohri et al., 2012), which generalizes the ramp loss also known as truncated hinge loss (Collobert et al., 2006a,b; Wu and Liu, 2007; Huang et al., 2014). The bounded ρ-margin loss function is a robust variant of the hinge loss used in standard SVMs (Cortes and Vapnik, 1995). Collobert et al. (2006a) proposed to replace the hinge loss in SVMs with the ramp loss to improve the scaling of SVMs in the sense that the number of support vectors is reduced. They also showed advantages of using the ramp loss for transductive SVMs (Collobert et al., 2006a,b). Wu and Liu (2007) and Brooks (2011) also investigated SVMs with ramp loss. Because of the increased robustness against outliers, Wu and Liu (2007) refer to their approach as *robust truncated hinge loss SVM* (RSVM). Recently, Huang et al. (2014) have proposed the *ramp loss linear programming SVM* (ramp- LPSVM).

In the following, we look at a basic generalization bound considering the ρ-margin loss function presented in the textbook by Mohri et al. (2012) and prove, inspired by the work of Ben-David et al. (2003a), that minimizing this margin bound is NP-hard for linear hypotheses. This implies that minimizing the ramp loss is NP-hard. The next section states the main result, which is proven in section 3 and quantitatively discussed in section 4.

2 MAIN RESULT

We start by defining our margin based surrogate loss. For any *target margin* $\rho > 0$, let the auxiliary function $\Phi_\rho : \mathbb{R} \to [0, 1]$ be given by

$$\Phi_\rho(x) := \begin{cases} 0 & \text{if } \rho \leq x \\ 1 - x/\rho & \text{if } 0 < x < \rho \\ 1 & \text{if } x \leq 0 \end{cases} .$$

The *ρ-margin loss function* $L_\rho : \mathbb{R} \times \mathbb{R} \to [0, 1]$ is now defined as $L_\rho(y, y') := \Phi_\rho(yy')$ for any $y, y' \in \mathbb{R}$ (Mohri et al., 2012), see Figure 1. Accordingly, the empirical *margin error* for $\rho > 0$ w.r.t. a sample S is given by $R_S^\rho(h) = \frac{1}{\ell} \sum_{i=1}^{\ell} L_\rho(y_i, h(\boldsymbol{x}_i))$.

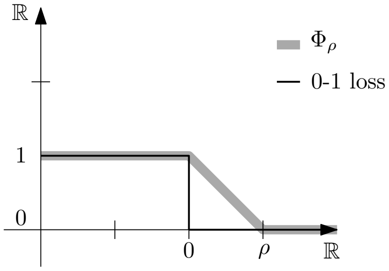


Figure 1: Plot of 0-1 loss $L_{0\text{-}1}(y, y')$ and Φ_ρ over yy'.

For $\rho \leq 1$ the ρ-margin loss function is a lower bound on the hinge loss $L_{\mathrm{hinge}}(y, y') = \max(0, 1 - yy')$ underlying support vector machines. The *ramp loss* (Collobert et al., 2006a,b; Huang et al., 2014), also called truncated hinge loss (Wu and Liu, 2007),[1] is defined as $L_{\mathrm{ramp}}(y, y') =$

[1] Other notions of ramp loss exist, however, our definition is arguably the most popular one, and the alternative definitions we know of can be mapped to the ρ-margin loss function in a similar way.

$\min(1, L_{\text{hinge}}(y, y'))$. It holds $L_{\text{ramp}} = L_1$, that is, for $\rho = 1$ the ρ-margin loss function equals the ramp loss.

We consider the following basic result, which corresponds to Corollary 5.1 by Mohri et al. (2012) for the special case of a linear kernel:

Theorem 2.1 (Corollary 5.1, Mohri et al., 2012)**.** *For any fixed $\lambda > 0$ and any fixed $\rho > 0$, the following holds for any $\delta > 0$ with probability at least $1 - \delta$ for any $h \in H_\lambda = \{\boldsymbol{x} \mapsto \langle \boldsymbol{w}, \boldsymbol{x} \rangle \mid \|\boldsymbol{w}\| \leq \lambda\}$:*

$$R^{0\text{-}1}(h) \leq R_S^\rho(h) + \frac{2\lambda}{\ell\rho} r + 3\sqrt{\frac{\log \frac{2}{\delta}}{2\ell}} \ ,$$

where $r = \sqrt{\sum_{i=1}^{\ell} \langle \boldsymbol{x}_i, \boldsymbol{x}_i \rangle}$.

We show that minimizing this upper bound on the generalization error is hard, that is, we show that the following problem is NP-hard:

Definition 2.2 (Minimization of Margin Bound with Linear Hypothesis (Min-MB-Linear))**.** Let $\lambda > 0$, let $\rho > 0$, let $H_\lambda = \{\boldsymbol{x} \mapsto \langle \boldsymbol{w}, \boldsymbol{x} \rangle \mid \|\boldsymbol{w}\| \leq \lambda\}$, and let S be a sample with ℓ elements. We formulate the problem as: Given $\varepsilon > 0$, does there exist $h^* \in H_\lambda$ such that

$$R_S^\rho(h^*) + \frac{2\lambda}{\ell\rho} r \leq \varepsilon \ ?$$

For the moment, we omit $3\sqrt{\frac{\log \frac{2}{\delta}}{2\ell}}$ in the problem formulation, since the choice of h^* is independent of this term (see section 4). Given definition 2.2, our main result reads:

Theorem 2.3. *For any $\rho > 0$ the Min-MB-Linear problem is NP-hard.*

We will proof the theorem by reducing the Max-E2-Sat problem to Min-MB-Linear. Given a value of ρ and a Max-E2-Sat instance, we will fix a λ depending on ρ and the number of variabels in the instance. Given ρ and λ, the $\frac{2\lambda}{\ell\rho} r$ term in Definition 2.2 is constant, and thus from Theorem 2.3 it follows:

Corollary 2.4. Let $H_\lambda = \{\boldsymbol{x} \mapsto \langle \boldsymbol{w}, \boldsymbol{x} \rangle \mid \|\boldsymbol{w}\| \leq \lambda\}$, and let S be a sample with ℓ elements. Then the following problem is NP-hard: Given $\epsilon > 0$, does there exist $h^* \in H_\lambda$ such that

$$R_S^\rho(h^*) \leq \epsilon \ ?$$

That is, minimizing the ramp loss is NP-hard. This basically follows from its similarity to the 0-1 loss, and we can adopt proof techniques that were applied to the 0-1 loss to show our results.

3 REDUCTION OF MAX-E2-SAT TO MIN-MB-LINEAR

To prove Theorem 2.3, we will reduce the Max-E2-Sat problem to the Min-MB-Linear problem. The construction and method of proof extends the methods used by Ben-David et al. (2003a). After stating Max-E2-Sat, we proceed by defining an instance transformation and a solution transformation. Then we proof the correctness of the reduction.

3.1 Max-E2-Sat

The problem we will reduce is the well-known satisfiability problem Max-E2-Sat defined as:

Definition 3.1 (Max-E2-Sat)**.** Given a set K of m clauses, where each clause is a disjunction of exactly two Boolean literals over a set of n variables, that is, $K = \{\alpha_1 \vee \beta_1, \ldots, \alpha_m \vee \beta_m\}$ where $\alpha_i \in V$ or $\overline{\alpha_i} \in V$, and $\beta_i \in V$ or $\overline{\beta_i} \in V$ for all $1 \leq i \leq m$ with variable set $V = \{\nu_1, \ldots, \nu_n\}$. Given $\varepsilon > 0$, does there exist $\boldsymbol{B} \in \{0, 1\}^n$ (the i'th component of $\boldsymbol{B}$ defines a truth assignment for ν_i, where 0 is interpreted as `false` and 1 as `true`) such that

$$R_K^{0\text{-}1}(\boldsymbol{B}) \leq \varepsilon \ ,$$

where $R_K^{0\text{-}1}(\boldsymbol{B}) = \frac{1}{m} \sum_{i=1}^{m} L_{0\text{-}1}(\boldsymbol{B}(\alpha_i \vee \beta_i), 1)$ and $\boldsymbol{B}(\alpha_i \vee \beta_i)$ is the evaluation of $\alpha_i \vee \beta_i$ using $\boldsymbol{B}$ as truth assignment?

Our results are based on the following theorem:

Theorem 3.2. *The Max-E2-Sat problem is NP-hard.*

A proof of this result can be found in the work of Garey et al. (1974).

3.2 The instance transformation

Let K be a set of m clauses over n variables (in the construction we will have $\dim(\mathcal{X}) = n + 1$). We fix the error of the Min-MB-Linear problem to be at most $\varepsilon' = \frac{2}{7} + \frac{\varepsilon}{7} + \frac{2\sqrt{\frac{n}{2} + 9\rho^2}}{\sqrt{7m}\rho}\sqrt{\frac{80\rho^2}{7} + 1}$ and fix $\lambda = \sqrt{\frac{n}{2} + 9\rho^2}$. For each clause $C \in K$ and each variable $\nu_j \in V$ we define

$$\psi_j(C) = \begin{cases} 2\sqrt{2}\rho & \text{if } \nu_j \in C \\ -2\sqrt{2}\rho & \text{if } \overline{\nu_j} \in C \\ 0 & \text{otherwise} \end{cases} \ .$$

Table 1: For ν_j and ν_k in C with $j \leq k$, we define $\varphi(C)$ for each clause $C \in K$ to consist of these seven samples $(u, 1)$, $(u_1, 1)$, $(u_2, 1)$, $(v_1, -1)$, $(v_2, -1)$, $(\hat{v_1}, -1)$, and $(\hat{v_2}, -1)$.

$$\begin{array}{rrrlclcl}
 & & & & \text{the } j\text{'th place} & & \text{the } k\text{'th place} & \\
\text{Labelled} & 1 & \left\{\begin{array}{r} u \;=\; \\ u_1 = \\ u_2 = \end{array}\right. & \begin{array}{l} (0,\ldots,0, \\ (0,\ldots,0, \\ (0,\ldots,0, \end{array} & \begin{array}{r} \psi_j(C) \\ -\psi_j(C) \\ \psi_j(C) \end{array} & \begin{array}{l} ,0,\ldots,0, \\ ,0,\ldots,0, \\ ,0,\ldots,0, \end{array} & \begin{array}{r} \psi_k(C) \\ \psi_k(C) \\ -\psi_k(C) \end{array} & \begin{array}{l} ,0,\ldots,0,1) \\ ,0,\ldots,0,1) \\ ,0,\ldots,0,1) \end{array} \\
 & & & & & & & \\
\text{Labelled} & -1 & \left\{\begin{array}{r} v_1 = \\ v_2 = \\ \hat{v_1} = \\ \hat{v_2} = \end{array}\right. & \begin{array}{l} (0,\ldots,0, \\ (0,\ldots,0, \\ (0,\ldots,0, \\ (0,\ldots,0, \end{array} & \begin{array}{r} 0 \\ \psi_j(C) \\ 0 \\ -\psi_j(C) \end{array} & \begin{array}{l} ,0,\ldots,0, \\ ,0,\ldots,0, \\ ,0,\ldots,0, \\ ,0,\ldots,0, \end{array} & \begin{array}{r} \psi_k(C) \\ 0 \\ -\psi_k(C) \\ 0 \end{array} & \begin{array}{l} ,0,\ldots,0,1) \\ ,0,\ldots,0,1) \\ ,0,\ldots,0,1) \\ ,0,\ldots,0,1) \end{array}
\end{array}$$

Assuming ν_j and ν_k occur in C with $j \leq k$, we define $\varphi(C)$ for each clause $C \in K$ to consist of the following seven samples $(u, 1)$, $(u_1, 1)$, $(u_2, 1)$, $(v_1, -1)$, $(v_2, -1)$, $(\hat{v_1}, -1)$, and $(\hat{v_2}, -1)$ constructed as shown in Table 1. Furthermore, we define $f(K) := \bigcup_{C \in K} \varphi(C)$ as the sample in the Min-MB-Linear problem. Note that by definition the number of examples in $f(K)$ is $\ell := 7m$, and since for a point $\boldsymbol{x}_1$ labelled 1 in $f(K)$ we have $\|\boldsymbol{x}_1\|^2 = 2(2\sqrt{2}\rho)^2 + 1 = 16\rho^2 + 1$ and for a point $\boldsymbol{x}_0$ labelled -1 we have $\|\boldsymbol{x}_0\|^2 = (2\sqrt{2}\rho)^2 + 1 = 8\rho^2 + 1$, we get

$$\begin{aligned} r &= \sqrt{\sum_{C \in K} \sum_{\boldsymbol{x} \in \varphi(C)} \|\boldsymbol{x}\|^2} \\ &= \sqrt{(3 \cdot 16\rho^2 + 3 + 4 \cdot 8\rho^2 + 4)m} \\ &= \sqrt{(80\rho^2 + 7)m} \ . \end{aligned}$$

The idea behind this construction is to create a configuration that allows to choose a hyperplane separating the point $(2\sqrt{2}\rho, 2\sqrt{2}\rho)$ from the points $(0, 2\sqrt{2}\rho)$ and $(2\sqrt{2}\rho, 0)$ with a margin of ρ, see Figure 2.

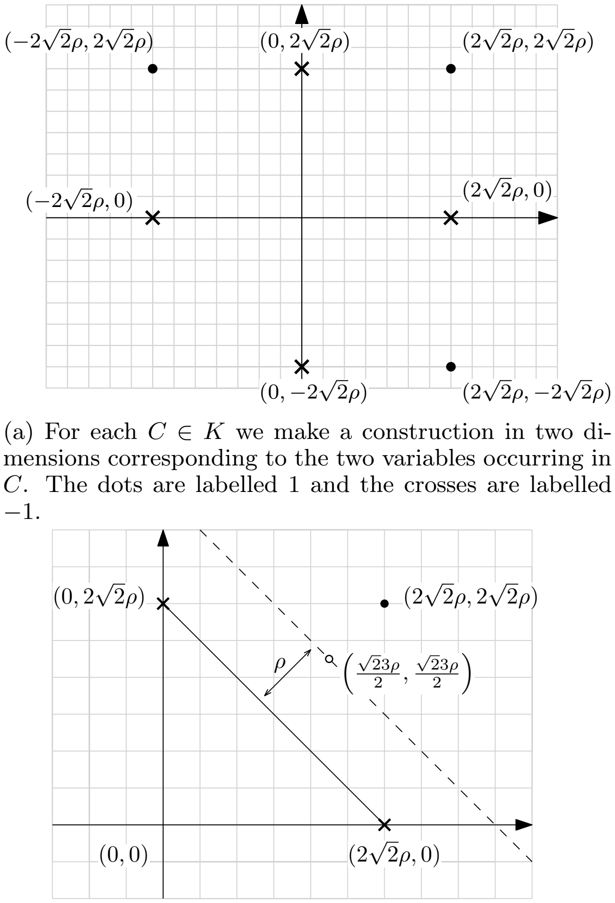
3.3 The solution transformation

Given a hypothesis $h \in H_\lambda$ defined by $h(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle$ for $\boldsymbol{w} = (w_1, \ldots, w_n, b) \in \mathbb{R}^{n+1}$ as solution to the Min-MB-Linear problem, we define a solution $g(h)$ to the Max-E2-Sat problem by $g(h) := (\xi_1(h), \ldots, \xi_n(h))^T$ with

$$\xi_i(h) := \begin{cases} 0 & \text{if } w_i < 0 \\ 1 & \text{if } w_i \geq 0 \end{cases} .$$

3.4 Analysis of the reduction

To prove the Theorem 2.3 we need the following lemma:


$(-2\sqrt{2}\rho, 2\sqrt{2}\rho)$
$(0, 2\sqrt{2}\rho)$
$(2\sqrt{2}\rho, 2\sqrt{2}\rho)$
$(-2\sqrt{2}\rho, 0)$
$(2\sqrt{2}\rho, 0)$
$(0, -2\sqrt{2}\rho)$
$(2\sqrt{2}\rho, -2\sqrt{2}\rho)$
$(0, 2\sqrt{2}\rho)$
$(2\sqrt{2}\rho, 2\sqrt{2}\rho)$
ρ
$\left(\frac{\sqrt{2}3\rho}{2}, \frac{\sqrt{2}3\rho}{2}\right)$
$(0, 0)$
$(2\sqrt{2}\rho, 0)$

(a) For each $C \in K$ we make a construction in two dimensions corresponding to the two variables occurring in C. The dots are labelled 1 and the crosses are labelled -1.

(b) This figure depicts the right upper quadrant of the construction of $\varphi(C)$. An optimal hyperplane illustrated by the dashed line, could pass through the point $(\frac{\sqrt{2}3\rho}{2}, \frac{\sqrt{2}3\rho}{2})$ with a margin of ρ.

Figure 2: The construction in the instance transformation.

Lemma 3.3. Let $\boldsymbol{B} \in \{0, 1\}^n$ and define h such that $g(h) = \boldsymbol{B}$ with $h(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle$ and $\boldsymbol{w} =$

$(w_1, \dots, w_n, b) \in \mathbb{R}^{n+1}$. Then the following holds:

$$R^{\rho}_{f(K)}(h) \geq \frac{2}{7} + \frac{1}{7} R^{\text{0-1}}_{K}(\boldsymbol{B})$$

Proof. Assume $C \in K$ is not satisfied by $\boldsymbol{B}$ and assume without loss of generality that $C = \nu_1 \vee \nu_2$. Hence, by definition we have $\psi_1(C) = 2\sqrt{2}\rho$ and $\psi_2(C) = 2\sqrt{2}\rho$. Since C is not satisfied by $\boldsymbol{B}$ this gives $B_1 = 0$ and $B_2 = 0$ and thus $w_1 < 0$ and $w_2 < 0$. Define $h^+ := \{\boldsymbol{x} \mid \langle \boldsymbol{w}, \boldsymbol{x} \rangle \geq 0\}$. Now we prove the following claim: The number of points in $\varphi(C)$ correctly labelled by h is at most four. In other words $7 - 7R^{\text{0-1}}_{\varphi(C)}(h) \leq 4$, which gives $3 \leq 7R^{\text{0-1}}_{\varphi(C)}(h)$. To prove this, we observe that the following holds:

1. If $(2\sqrt{2}\rho, 2\sqrt{2}\rho, 0, \dots, 0, 1) \in h^+$ (i.e. $\langle \boldsymbol{w}, (2\sqrt{2}\rho, 2\sqrt{2}\rho, 0, \dots, 0, 1) \rangle \geq 0$) then $w_1 2\sqrt{2}\rho + w_2 2\sqrt{2}\rho + b \geq 0$, hence $w_1 2\sqrt{2}\rho + w_2 2\sqrt{2}\rho \geq -b$. However, since $w_1 < 0$ and $w_2 < 0$ we have $w_1 2\sqrt{2}\rho \geq -b$ and $w_2 2\sqrt{2}\rho \geq -b$, so $(2\sqrt{2}\rho, 0, 0, \dots, 0, 1) \in h^+$ and $(0, 2\sqrt{2}\rho, 0, \dots, 0, 1) \in h^+$.
2. If $(2\sqrt{2}\rho, -2\sqrt{2}\rho, 0, \dots, 0, 1) \in h^+$ then $w_1 2\sqrt{2}\rho - w_2 2\sqrt{2}\rho \geq -b$, hence $-w_2 2\sqrt{2}\rho \geq -b$ so $(0, -2\sqrt{2}\rho, 0, \dots, 0, 1) \in h^+$.
3. If $(-2\sqrt{2}\rho, 2\sqrt{2}\rho, 0, \dots, 0, 1) \in h^+$ then $-w_1 2\sqrt{2}\rho + w_2 2\sqrt{2}\rho \geq -b$, hence $-w_1 2\sqrt{2}\rho \geq -b$ so $(-2\sqrt{2}\rho, 0, 0, \dots, 0, 1) \in h^+$.

As $(2\sqrt{2}\rho, 2\sqrt{2}\rho, 0, \dots, 0)$, $(2\sqrt{2}\rho, -2\sqrt{2}\rho, 0, \dots, 0)$, and $(-2\sqrt{2}\rho, 2\sqrt{2}\rho, 0, \dots, 0)$ are all labelled 1 in the definition of $\varphi(C)$, and $(2\sqrt{2}\rho, 0, 0, \dots, 0)$, $(0, 2\sqrt{2}\rho, 0, \dots, 0)$, $(0, -2\sqrt{2}\rho, 0, \dots, 0)$, and $(-2\sqrt{2}\rho, 0, 0, \dots, 0)$ are labelled -1 in $\varphi(C)$ we have that h^+ contains at least as many points from $\varphi(C)$ labelled -1 as points labelled 1.

Now, define $\sharp h^+_{-1}$ to be the number of points from $\varphi(C)$ labelled -1 that are contained in h^+ and $\sharp h^+_1$ to be the number of points from $\varphi(C)$ labelled 1 that are contained in h^+. From the above we have $\sharp h^+_1 \leq \sharp h^+_{-1}$, hence $4 - \sharp h^+_{-1} \leq 4 - \sharp h^+_1$. Now, since $\varphi(C)$ contains four points labelled -1, the number of points labelled -1 that h labels correctly is $4 - \sharp h^+_{-1}$. Similarly, $\sharp h^+_1$ is the number of points labelled 1 that h labels correctly, hence the total number of points that h labels correctly satisfies $7 - 7R^{\text{0-1}}_{\varphi(C)}(h) = 4 - \sharp h^+_{-1} + \sharp h^+_1 \leq 4$ as wanted.

On the other hand, we assume $C \in K$ is satisfied by $\boldsymbol{B}$ and assume again without loss of generality that $C = \nu_1 \vee \nu_2$. Since $\boldsymbol{B}$ satisfies C, at least one of w_1 or w_2 is greater than or equal to 0. Similarly as before we prove the claim: The number of points in $\varphi(C)$ correctly labelled by h is at most five. In other words $7 - 7R^{\text{0-1}}_{\varphi(C')}(h) \leq 5$, which gives $2 \leq 7R^{\text{0-1}}_{\varphi(C')}(h)$. To prove this, we define h^+ as before and observe that the following holds:

1. If $(2\sqrt{2}\rho, -2\sqrt{2}\rho, 0, \dots, 0, 1) \in h^+$ and $(-2\sqrt{2}\rho, 2\sqrt{2}\rho, 0, \dots, 0, 1) \in h^+$ then $w_1 2\sqrt{2}\rho - w_2 2\sqrt{2}\rho + b \geq 0$ and $-w_1 2\sqrt{2}\rho + w_2 2\sqrt{2}\rho + b \geq 0$. Adding these two inequalities gives $2b \geq 0$, hence $b \geq 0$. Since $w_1 \geq 0$ or $w_2 \geq 0$ we have $w_1 2\sqrt{2}\rho + b \geq 0$ or $w_2 2\sqrt{2}\rho + b \geq 0$, hence $(2\sqrt{2}\rho, 0, 0, \dots, 0, 1) \in h^+$ or $(0, 2\sqrt{2}\rho, 0, \dots, 0, 1) \in h^+$.
2. If $(2\sqrt{2}\rho, 2\sqrt{2}\rho, 0, \dots, 0, 1) \in h^+$ and $(-2\sqrt{2}\rho, 2\sqrt{2}\rho, 0, \dots, 0, 1) \in h^+$ then $w_1 2\sqrt{2}\rho + w_2 2\sqrt{2}\rho + b \geq 0$ and $-w_1 2\sqrt{2}\rho + w_2 2\sqrt{2}\rho + b \geq 0$. Adding these two inequalities gives $2w_2 2\sqrt{2}\rho + 2b \geq 0$, hence $w_2 2\sqrt{2}\rho + b \geq 0$ so $(0, 2\sqrt{2}\rho, 0, \dots, 0, 1) \in h^+$.
3. If $(2\sqrt{2}\rho, 2\sqrt{2}\rho, 0, \dots, 0, 1) \in h^+$ and $(2\sqrt{2}\rho, -2\sqrt{2}\rho, 0, \dots, 0, 1) \in h^+$ then $w_1 2\sqrt{2}\rho + w_2 2\sqrt{2}\rho + b \geq 0$ and $w_1 2\sqrt{2}\rho - w_2 2\sqrt{2}\rho + b \geq 0$. Again, adding these two gives $2w_1 2\sqrt{2}\rho + 2b \geq 0$ so $w_1 2\sqrt{2}\rho + b \geq 0$, hence $(2\sqrt{2}\rho, 0, 0, \dots, 0, 1) \in h^+$.

The above gives that if h^+ contains two examples labelled 1 then h^+ contains at least one example labelled -1. Furthermore, it gives that if h^+ contains all three examples labelled 1 then h^+ contains at least two examples labelled -1. By defining $\sharp h^+_{-1}$ and $\sharp h^+_1$ as before, this provides us with the inequality $\sharp h^+_1 \leq \sharp h^+_{-1} + 1$, hence $4 - \sharp h^+_{-1} \leq 5 - \sharp h^+_1$ which again gives $7 - 7R^{\text{0-1}}_{\varphi(C')}(h) = 4 - \sharp h^+_{-1} + \sharp h^+_1 \leq 5$ proving the claim.

By defining m_1 to be the number of clauses satisfied by $\boldsymbol{B}$, $m_0 = m - m_1$ to be the number of clauses not satisfied by $\boldsymbol{B}$, and

$$f_1(K) := \bigcup_{\substack{C \in K \\ \boldsymbol{B}(C)=1}} \varphi(C) \text{ and } f_0(K) := \bigcup_{\substack{C \in K \\ \boldsymbol{B}(C)=0}} ,$$

where $\boldsymbol{B}(C) = 1$ if C is satisfied by $\boldsymbol{B}$ and $\boldsymbol{B}(C) = 0$ otherwise, we get by the above claims:

$$7m_1 R^{\text{0-1}}_{f_1(K)}(h) = 7m_1 \frac{1}{7m_1} \sum_{\substack{C \in K \\ \boldsymbol{B}(C)=1}} 7R^{\text{0-1}}_{\varphi(C)}(h)$$

$$= 7 \sum_{\substack{C \in K \\ \boldsymbol{B}(C)=1}} R^{0\text{-}1}_{\varphi(C)}(h) \geq 7\frac{2}{7}m_1 = 2m_1$$

$$7m_0 R^{0\text{-}1}_{f_0(K)}(h) = 7m_0 \frac{1}{7m_0} \sum_{\substack{C \in K \\ \boldsymbol{B}(C)=0}} 7R^{0\text{-}1}_{\varphi(C)}(h)$$

$$= 7 \sum_{\substack{C \in K \\ \boldsymbol{B}(C)=0}} R^{0\text{-}1}_{\varphi(C)}(h) \geq 7\frac{3}{7}m_0 = 3m_0$$

Since $f(K) = f_1(K) \cup f_0(K)$ (because $f_1(K)$ and $f_0(K)$ are clearly disjoint) we get from these inequalities

$$\begin{aligned}
\ell R^{0\text{-}1}_{f(K)}(h) &= 7m_1 R^{0\text{-}1}_{f_1(K)}(h) + 7m_0 R^{0\text{-}1}_{f_0(K)}(h) \\
&\geq 2m_1 + 3m_0 \\
&= 2m_1 + 3(m - m_1) = 3m - m_1 \\
&= 3m - m + mR^{0\text{-}1}_K(\boldsymbol{B}) \\
&= 2m + mR^{0\text{-}1}_K(\boldsymbol{B}) \ .
\end{aligned}$$

Inserting $m = \frac{\ell}{7}$ we get $R^{0\text{-}1}_{f(K)}(h) \geq \frac{2}{7} + \frac{1}{7}R^{0\text{-}1}_K(\boldsymbol{B})$, hence by the inequality $R^{\rho}_{f(K)}(h) \geq R^{0\text{-}1}_{f(K)}(h)$ the result follows. □

This allows us to prove Theorem 2.3.

Proof of Theorem 2.3. First, we assume that the answer to the Max-E2-Sat problem is "yes", that is, we assume that there exists $\boldsymbol{B} = (B_1, \dots, B_n) \in \{0,1\}^n$ such that $R^{0\text{-}1}_K(\boldsymbol{B}) \leq \varepsilon$. We define $h(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle$ where $\boldsymbol{w} = (w_1, \dots, w_n, b)$ is given by

$$w_i = \begin{cases} \frac{1}{\sqrt{2}} & \text{, if } B_i = 1 \\ -\frac{1}{\sqrt{2}} & \text{, if } B_i = 0 \end{cases}$$

and $b = -3\rho$. From this we see that $\|\boldsymbol{w}\| = \sqrt{w_1^2 + \cdots + w_n^2 + b^2} = \sqrt{\frac{n}{2} + 9\rho^2} = \lambda$, hence $h \in H_\lambda$.

Let $C \in K$. From this, we note that no point labelled -1 in $\varphi(C)$ is in the set $h^+ := \{\boldsymbol{x} \mid \langle \boldsymbol{w}, \boldsymbol{x} \rangle \geq 0\}$ since a point $\boldsymbol{x}_0$ labelled -1 has only one non-zero component (apart from the constant 1 in the end of each example) which is either $2\sqrt{2}\rho$ or $-2\sqrt{2}\rho$, hence we have

$$\langle \boldsymbol{w}, \boldsymbol{x}_0 \rangle \leq \frac{2\sqrt{2}\rho}{\sqrt{2}} - 3\rho = -\rho < 0 \ .$$

This implies that h classifies all points labelled -1 correctly. A point $\boldsymbol{x}_1$ labelled 1 in $\varphi(C)$ has two non-zero components (again apart from the constant 1), therefore, if both of these components have the same sign as the corresponding components of $\boldsymbol{w}$, we have

$$\langle \boldsymbol{w}, \boldsymbol{x}_1 \rangle = \frac{4\sqrt{2}\rho}{\sqrt{2}} - 3\rho = \rho > 0 \ .$$

On the other hand, if the signs of the components differ we have

$$\langle \boldsymbol{w}, \boldsymbol{x}_1 \rangle \leq b = -3\rho < 0 \ ,$$

which gives that $\boldsymbol{x}_1 \in h^+$ if and only if both of the non-zero components (not 1) have the same sign as those of $\boldsymbol{w}$. Now, if $\boldsymbol{B}$ satisfies C such a point must exist by definition of $\varphi(C)$, since if we assume without loss of generality that $C = \nu_1 \vee \nu_2$ then $B_1 = 1$ or $B_2 = 1$ (or both), hence at least one of the first two components of $\boldsymbol{w}$ is $2\sqrt{2}\rho$. As $\varphi(C)$ contains $u = (2\sqrt{2}\rho, 2\sqrt{2}\rho, 0, \dots, 0)$, $u_1 = (-2\sqrt{2}\rho, 2\sqrt{2}\rho, 0, \dots, 0)$, and $u_2 = (2\sqrt{2}\rho, -2\sqrt{2}\rho, 0, \dots, 0)$, one of these must have the same sign in the first two components as $\boldsymbol{w}$. Therefore, for each clause C' satisfied by $\boldsymbol{B}$ we have that h correctly labels at least five of the examples in $\varphi(C')$ (four labelled -1 and one labelled 1), and for each clause C'' not satisfied by $\boldsymbol{B}$, h correctly labels at least four of the examples in $\varphi(C'')$. Now, define m_1 to be the number of clauses satisfied by $\boldsymbol{B}$ and let $m_0 = m - m_1$, that is m_0 is the number of clauses not satisfied by $\boldsymbol{B}$. This gives $\ell - \ell R^{0\text{-}1}_{f(K)}(h) \geq 5m_1 + 4m_0 = m_1 + 4m = m - mR^{0\text{-}1}_K(\boldsymbol{B}) + 4m$.

Now, we claim that $\ell - \ell R^{0\text{-}1}_{f(K)}(h) = \ell - \ell R^{\rho}_{f(K)}(h)$, or equivalently $R^{0\text{-}1}_{f(K)}(h) = R^{\rho}_{f(K)}(h)$. This is again equivalent to stating that no point $\boldsymbol{x} \in \varphi(C)$ for any $C \in K$ satisfies $|\langle \boldsymbol{w}, \boldsymbol{x} \rangle| < \rho$, since the margin loss outside the strip $|\langle \boldsymbol{w}, \boldsymbol{x} \rangle| < \rho$ is equal to the 0-1 loss. We can easily verify this by calculating $|\langle \boldsymbol{w}, \boldsymbol{x} \rangle|$ for every point in $\varphi(C)$ as shown in Table 2, where we only calculate the two dimensions in which the components (except the bias) of $\boldsymbol{x}$ are non-zero.

This gives $\ell - \ell R^{\rho}_{f(K)}(h) \geq m - mR^{0\text{-}1}_K(\boldsymbol{B}) + 4m$, hence by $m = \frac{\ell}{7}$ we have

$$\ell R^{\rho}_{f(K)}(h) \leq \ell - \frac{5}{7}\ell + \frac{\ell}{7}R^{0\text{-}1}_K(\boldsymbol{B}) \ \Rightarrow$$

$$R^{\rho}_{f(K)}(h) \leq \frac{2}{7} + \frac{1}{7}R^{0\text{-}1}_K(\boldsymbol{B}) \ .$$

Using the definition of λ and the calculated value of r we get

$$R^{\rho}_{f(K)}(h) + \frac{2\lambda}{\ell\rho}r \leq \frac{2}{7} + \frac{1}{7}R^{0\text{-}1}_K(\boldsymbol{B})$$

Table 2: This table displays the values of $|\langle \boldsymbol{w}, \boldsymbol{x}\rangle|$ in the 2-dimensional case (where $a := 2\sqrt{2}\rho$). The dashed vertical line indicates the separation of the points labelled 1 and -1, and each value of $\boldsymbol{w}$ corresponds to a hyperplane.

		Values of $\boldsymbol{x}$						
		(a,a)	$(-a,a)$	$(a,-a)$	$(a,0)$	$(0,a)$	$(-a,0)$	$(0,-a)$
Values of $\boldsymbol{w}$	$(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$	ρ	3ρ	3ρ	ρ	ρ	5ρ	5ρ
	$(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$	3ρ	ρ	7ρ	5ρ	ρ	ρ	5ρ
	$(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})$	3ρ	7ρ	ρ	ρ	5ρ	5ρ	ρ
	$(-\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})$	7ρ	3ρ	3ρ	5ρ	5ρ	ρ	ρ

$$+\frac{2\sqrt{\frac{n}{2}+9\rho^2}}{\ell\rho}\sqrt{(80\rho^2+7)m}$$

$$\le\frac{2}{7}+\frac{\varepsilon}{7}+\frac{2\sqrt{\frac{n}{2}+9\rho^2}}{\sqrt{7m}\rho}\sqrt{\frac{80\rho^2}{7}+1}$$

$$=\varepsilon'$$

as wanted.

On the other hand, we assume that the answer to the Min-MB-Linear problem is "yes", that is, we assume that there exists $h \in H$ such that $R^{\rho}_{f(K)}(h) + \frac{2\lambda}{\ell\rho}r \le \varepsilon'$, hence by definition of λ and r we assume $R^{\rho}_{f(K)}(h) \le \frac{2}{7} + \frac{\varepsilon}{7}$.

Now, we want to show that $\boldsymbol{B}$ constructed from h as in the solution transformation satisfies $R^{0\text{-}1}_{K}(\boldsymbol{B}) \le \varepsilon$. To achieve this we observe by Lemma 3.3 that we have

$$R^{0\text{-}1}_{K}(\boldsymbol{B}) \le 7R^{\rho}_{f(K)}(h) - 2 \le 7\left(\frac{2}{7}+\frac{\varepsilon}{7}\right) - 2 = \varepsilon$$

as wanted, which completes the proof. □

4 QUANTITATIVE ANALYSIS

In this section, we will discuss our result quantitatively. The term bounded in the Min-MB-Linear problem is an upper bound on the generalization error using the 0-1 loss with the term $3\sqrt{\frac{\log\frac{2}{\delta}}{2\ell}}$ added. For this term to make sense from the machine learning point of view, it must be less than 1 since otherwise the bound is trivial. In the reduction to the Min-MB-Linear problem we ask whether we can find a hypothesis h such that

$$R^{\rho}_{S}(h) + \frac{2\lambda}{\ell\rho}r \le \frac{2}{7}+\frac{\varepsilon}{7}+\frac{2\sqrt{\frac{n}{2}+9\rho^2}}{\sqrt{7m}\rho}\sqrt{\frac{80\rho^2}{7}+1} \ .$$

Assuming $0 < \varepsilon \le 1$ (which makes sense in context of the Max-E2-Sat problem) we thus need

$$m > \frac{7}{25}\left(\frac{2\sqrt{\frac{n}{2}+9\rho^2}}{\rho}\sqrt{\frac{80\rho^2}{7}+1}+\frac{3}{\sqrt{2}}\sqrt{\log\frac{2}{\delta}}\right)^2 \ .$$

On the other hand, the maximal number of clauses we can make with n variables, without making duplicates, is $4\binom{n}{2} = 4\frac{n!}{2!(n-2)!} = 2n(n-1)$, since there are $\binom{n}{2}$ possible choices for picking two variables and each pair of variables comes in four different versions when taking negation into account. This gives $m \le 2n(n-1)$, hence we need

$$2n(n-1) >$$
$$\frac{7}{25}\left(\frac{2\sqrt{\frac{n}{2}+9\rho^2}}{\rho}\sqrt{\frac{80\rho^2}{7}+1}+\frac{3}{\sqrt{2}}\sqrt{\log\frac{2}{\delta}}\right)^2 \ .$$

The left hand side scales quadratically, the right hand side linearly with n, see Figure 3 for illustrations. This shows that the upper bound is indeed informative from a machine learning point of view.

5 DISCUSSION AND CONCLUSION

The main result shows that minimizing a basic margin bound is NP-hard when considering linear hypotheses and the ρ-margin loss function. This directly implies NP-hardness of minimizing the ρ-margin loss and in particular the ramp loss, a loss function used in many learning algorithms (Collobert et al., 2006a,b; Wu and Liu, 2007; Brooks, 2011; Huang et al., 2014). This extends the well-known result that minimizing the 0-1 error is NP-hard (see Ben-David et al., 2003a). Unfortunately, our proof cannot easily be extended to not bounded loss functions such as the hinge loss.

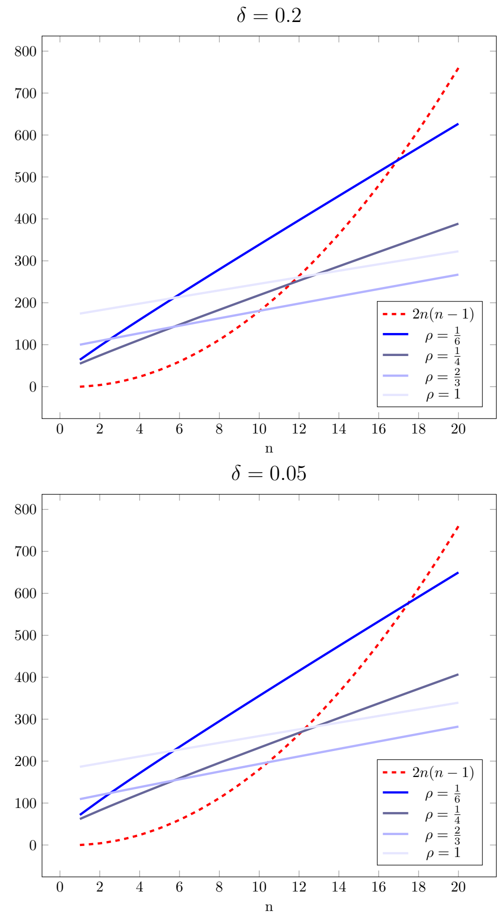


Figure 3: The functions $n \mapsto 2n(n-1)$ and $n \mapsto \frac{7}{25}\left(\frac{2\sqrt{\frac{n}{2}+9\rho^2}}{\rho}\sqrt{\frac{80\rho^2}{7}+1}+\frac{3}{\sqrt{2}}\sqrt{\log\frac{2}{\delta}}\right)^2$ for different values of ρ and δ. We observe that the value of $2n(n-1)$ is larger than the other functions for every $n \geq N$ for some N around 17. The first axis corresponds to the dimension of the input space and the second to the size of the sample.

This study is motivated by the analysis of generalization bounds from which model selection strategies are derived. We do not claim that optimizing the upper bound given by Theorem 2.1 is a proper model selection strategy, but many model selection methods are based on optimizing—arguably more sophisticated—generalization bounds. Our result indicates that the model selection of large margin classifiers may be a hard problem if based on generalization bounds, which would justify the use of heuristics for model selection such as gradient descent methods on multimodal objective functions (e.g., Chapelle et al., 2002; Glasmachers and Igel, 2010).

Theorem 2.3 does not directly extend to the nonlinear case involving non-linear kernels. Ben-David et al. (2003a) prove that while a specific sample may give a computational intractable problem in one class of hypotheses, it may be easily learned in another class. Thus, a construction proving hardness in one class of hypotheses does not necessarily carry over to another class of hypotheses. However, since the method for proving Theorem 2.3 is based on an extension of the method for proving that finding optimal half-spaces is hard, we hypothesize that a similar extension exists for results proving that learning kernel-based half-spaces is hard, see Shalev-Shwartz et al. (2010) for research in this direction.

Acknowledgements

Support from the European Commission through project AKMI (PCIG10-GA-2011-303655) is gratefully acknowledged.

References

P. Bartlett and S. Ben-David. Hardness results for neural networks approximation problems. *Theoretical Computer Science*, 284:53–66, 2002.

S. Ben-David, N. Eiron, and P. M. Long. On the difficulty of approximately maximizing agreements. *Journal of Computer and System Sciences*, 66(3):496–514, 2003a.

S. Ben-David, N. Eiron, and H. U. Simon. Limitations of learning via embeddings in Euclidean half spaces. *Journal of Machine Learning Research*, 3:441–461, 2003b.

S. Boucheron, O. Bousquet, and G. Lugosi. Theory of Classification: a Survey of Some Recent Advances. *ESAIM: Probability and Statistics*, 9: 323–375, 2005.

J. P. Brooks. Support vector machines with the ramp loss and the hard margin loss. *Operations Research*, 59(2):467–479, 2011.

O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46 (1):131–159, 2002.

R. Collobert, F. Sinz, J. Weston, and L. Bottou. Large scale transductive SVMs. *The Journal of Machine Learning Research*, 7:1687–1712, 2006a.

R. Collobert, F. Sinz, J. Weston, and L. Bottou. Trading convexity for scalability. In *Proceedings* *of the 23rd International Conference on Machine Learning (ICML)*, pages 201–208. ACM, 2006b.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

V. Feldman, V. Guruswami, P. Raghavendra, and Y. Wu. Agnostic learning of monomials by halfspaces is hard. In *Symposium on Foundations of Computer Science (FOCS)*, pages 385–394, 2009.

M. R. Garey, D. S. Johnson, and L. Stockmeyer. Some simplified np-complete problems. In *Proceedings of the Sixth Annual ACM Symposium on Theory of Computing (STOC)*, pages 47–63. ACM, 1974.

T. Glasmachers and C. Igel. Maximum likelihood model selection for 1-norm soft margin SVMs with multiple parameters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(8):1522–1528, 2010.

X. Huang, L. Shi, and J. A. K. Suykens. Ramp loss linear programming support vector machine. *Journal of Machine Learning Research*, 15:2185–2211, 2014.

D. R. Hush. Training a sigmoidal node is hard. *Neural Computation*, 11(5):1249–1260, 1999.

D. S. Johnson and F. P. Preparata. The densest hemisphere problem. *Theoretical Computer Science*, 6(1):93–107, 1978.

M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning.* MIT Press, 2012.

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

S. Shalev-Shwartz, O. Shamir, and K. Sridharan. Learning kernel-based halfspaces with the 0-1 loss. *SIAM Journal on Computing*, 40(6):1623–1646, 2010.

J. Šíma. Training a single sigmoidal neuron is hard. *Neural Computation*, 14(11):2709–2728, 2002.

V. Vapnik. *Statistical Learning Theory.* Wiley, 1998. ISBN 978-0-471-03003-4.

Y. Wu and Y. Liu. Robust truncated hinge loss support vector machines. *Journal of the American Statistical Association*, 102(479):974–983, 2007.