

Optimality of SVM: Novel Proofs and Tighter Bounds

Steve Hanneke

Princeton, NJ, USA

Aryeh Kontorovich

Ben-Gurion University, Beer Sheva, Israel

Abstract

We provide a new proof that the expected error rate of consistent support vector machines matches the minimax rate (up to a constant factor) in its dependence on the sample size and margin. The upper bound was originally established by [1], while the lower bound follows from an argument of [2] together with reasoning about the VC dimension of large-margin classifiers. Our proof differs from the original in that many of our steps concern reasoning about the primal space, while the original carried out these steps by reasoning about the dual space. Our approach provides a unified framework for analyzing both the homogeneous and non-homogeneous cases, with slightly better results for the former. The fact that our analysis explicitly handles the non-homogeneous case offers significant improvements in the bounds compared to the usual textbook approach of reducing to the homogeneous case. We also extend our proof to provide a new upper bound on the error rate of transductive SVM, which yields an improved constant factor compared to inductive SVM. In addition to these bounds on the expected error rate, we also provide a simple proof of a margin-based PAC-style bound for support vector machines, and an extension of the agnostic PAC analysis that explicitly handles the non-homogeneous case.

Keywords: Statistical learning theory; Support vector machine; PAC learning; Margin bound; Classification; Generalization bound

1. Introduction

Margin and VC-dimension based sample complexity bounds are a crown jewel of the PAC theory of supervised binary classification. In particular, a considerable amount of theory has been devoted to linear classifiers. The agnostic case is well-understood: if all but a few of m labeled data points residing on the n -dimensional unit sphere are linearly separated with margin at least γ (the few

Email addresses: steve.hanneke@gmail.com (Steve Hanneke), karyeh@cs.bgu.ac.il (Aryeh Kontorovich)

7 exceptions being treated as sample errors) then the expected excess risk decays
 8 as [3, 4] $\Theta\left(\sqrt{\min\{n, 1/\gamma^2\}/m}\right)$. For the separable case, in which there exists
 9 a hyperplane in \mathbb{R}^n consistent with the m sample points and having margin at
 10 least γ , it follows from known results that the best guarantee on the expected
 11 risk by any learning algorithm is lower-bounded by

$$\Omega(\min\{n, 1/\gamma^2\}/m). \quad (1)$$

12 Similarly, any generalization bound that holds with probability $1 - \delta$ is lower
 13 bounded by

$$\Omega\left(\left(\min\{n, 1/\gamma^2\} + \log(1/\delta)\right)/m\right). \quad (2)$$

14 For completeness, a proof sketch of the lower bound (1) is included in Ap-
 15 pendix Appendix C; it follows by combining a result of [2] with a simple shat-
 16 terability argument for $1/\gamma^2$ coordinate vectors by γ -margin separators. The
 17 proof of (2) follows from analogous arguments, except replacing the lower bound
 18 of [2] with that of [5].

19 The work of [1] establishes that the support vector machine indeed achieves
 20 an expected error rate guarantee of the form (1), as its expected error rate on
 21 m samples is at most a value proportional to

$$\mathbb{E}[\min\{n + 1, 1/\gamma_{m+1}^2\}]/(m + 1), \quad (3)$$

22 where γ_{m+1} is the maximum margin achievable by a linear separator for $m + 1$
 23 random data points (which is a random variable). Standard results in vari-
 24 ous textbooks (e.g., [6]) also state an upper bound on the error rate for the
 25 homogeneous support vector machine holding with probability $1 - \delta$:

$$O\left(\frac{1}{m} \left(\min\{n, 1/\gamma^2\} \log\left(\frac{m}{\min\{n + 1, 1/\gamma^2\}}\right) \log(m) + \log\left(\frac{1}{\delta}\right) \right)\right). \quad (4)$$

26 **Main results.** Our present work serves to fill some of the gaps in the known
 27 results. First, we sharpen the upper bound (4), removing a factor $\log(m)$ from
 28 the first term in this bound to get¹

$$O\left(\frac{1}{m} \left(\min\{n + 1, 1/\gamma^2\} \log\left(\frac{m}{\min\{n + 1, 1/\gamma^2\}}\right) + \log\left(\frac{1}{\delta}\right) \right)\right). \quad (5)$$

29 It remains a major open problem to determine whether the SVM achieves an
 30 upper bound matching (2) up to constant factors. This stronger guarantee is

¹We note that proofs of this type of refinement have been known in “folklore” form for some time. In particular, we thank John Shawe-Taylor [7] for sharing unpublished lecture notes on a technique for achieving this (via bounding the covering numbers). However, we also note that our proof is significantly simpler and leads to smaller constant factors, compared to these folklore proofs.

31 already known to hold for an algorithm based on the Perceptron learning rule.
32

33 Second, our proof of the bound (5) directly addresses the presence of the
34 *bias term* in the support vector machine. As is well known (both in prac-
35 tice and in theory), the non-homogeneous linear separation problem (allowing
36 a nonzero bias term) can be represented as a homogeneous linear separation
37 problem in one additional dimension (augmenting each example with an addi-
38 tional “dummy” feature, whose value is fixed to 1). The traditional approach to
39 analyzing large margin separators focuses on homogeneous separators, suppos-
40 ing that this transformation has already been applied. However, we note that
41 the value of the margin can change *dramatically* under this transformation.
42 The margin appearing in the bound is the *geometric* margin, which involves a
43 normalized weight vector. By transforming to the homogeneous case, we must
44 include the bias term in the vector being normalized, which can increase the
45 norm by an arbitrary amount. In contrast, the (non-homogeneous) support
46 vector machine maximizes the margin *without* including the bias term in the
47 normalization. The margins appearing in our results correspond to this latter
48 notion of margin, which therefore are better representations of the behavior of
49 the SVM.

50 Additionally, this work provides a new proof of the upper bound (3) on the
51 expected error rate of SVM. Unlike the existing proof of [1], our proof treats
52 both the homogeneous and non-homogeneous cases simultaneously, and in a
53 unified way (and without reduction to the homogeneous case). Furthermore, the
54 argument extends in a natural way to provide the first published bounds on the
55 expected error rate of *transductive* SVM matching the form (3) (again, for both
56 the homogeneous and non-homogeneous cases).³ The bounds for transductive
57 SVM offer improvements over those for inductive SVM in the constant factors.
58 In addition to these results for the realizable case, we also derive an agnostic PAC
59 bound relevant to SVM. As with our results for the realizable case, our agnostic
60 PAC bound differs from the standard treatment in that it explicitly accounts
61 for the bias term in non-homogeneous SVM, and this fact offers significant
62 quantifiable improvements in the bound, compared to the standard approach
63 of reducing to the homogeneous case. These results for the agnostic case are
64 presented in Section 8.

65 2. Definitions and Notation

66 We consistently use m to denote sample size, with $[m] := \{1, \dots, m\}$, and
67 $n \geq 2$ to denote the dimension of the Euclidean instance space. Vectors are
68 denoted in boldface ($\mathbf{x} = (x_1, \dots, x_n)$), and are capitalized when random. The

²In particular, Littlestone’s online-to-batch conversion [8] combined with Novikoff’s Perceptron mistake bound [9] yields the upper bound matching (2).

³We note, however, that one can modify the argument of [1] to obtain a similar result for transductive SVM (though again, that argument would only apply to non-homogeneous SVM).

69 standard inner product is denoted by $\langle \mathbf{x}, \mathbf{z} \rangle = \sum_{i=1}^n x_i z_i$, and induces the Euclidean norm $\|\mathbf{x}\|^2 = \langle \mathbf{x}, \mathbf{x} \rangle$. We write \mathbf{x}_i to mean the i^{th} vector in a sequence, and x_{ij} to denote its j^{th} component, should the need ever arise. Sequences $(\mathbf{x}_1, \dots, \mathbf{x}_m)$ will occasionally be abbreviated to \mathbf{x}_1^m . Set cardinalities are denoted by $\text{card}(\cdot)$ and $\mathbb{1}[\cdot]$ denotes the 0-1 truth value of the predicate inside the brackets. For $\boldsymbol{\alpha} \in \mathbb{R}^m$, its *support* is defined by $\text{supp}(\boldsymbol{\alpha}) = \{i \in [m] : \alpha_i \neq 0\}$ and $\|\boldsymbol{\alpha}\|_0 := \text{card}(\text{supp}(\boldsymbol{\alpha}))$. The nonnegative reals are denoted by $\mathbb{R}_+ := [0, \infty)$, the extended reals are denoted by $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$, and the Euclidean unit sphere is denoted by $\mathbb{S}^n = \{x \in \mathbb{R}^n : \|x\| = 1\}$. Additionally, for any $t \in \bar{\mathbb{R}}$, we denote $\text{sign}(t) = 2\mathbb{1}[t \geq 0] - 1$.

79 As per the standard consistent-PAC setting, $\mathbf{X}_1, \mathbf{X}_2, \dots$ will be an i.i.d. sequence of data points, drawn from an arbitrary fixed distribution on \mathbb{R}^n , and labeled by a target hyperplane. Throughout the paper, a dataset⁴ \mathcal{D} is always understood to contain example-label pairs $(\mathbf{x}, y) \in \mathbb{R}^n \times \{-1, 1\}$, either randomly generated (when capitalized) or else arbitrarily chosen, and will always be assumed to be *strictly linearly separable* (in a sense defined below), except in the results on agnostic learning. To indicate that the i^{th} data point has been omitted, we will write $\mathcal{D}_{-i} := \mathcal{D} \setminus \{(\mathbf{x}_i, y_i)\}$. All probabilities and expectations will be with respect to the fixed distribution or its appropriate k -fold products, as will be clear from context.

89 **Homogeneous Case vs Non-homogeneous Case.** The support vector machine can be formulated in two distinct ways, depending on whether we allow a *bias* term. Specifically, in the *homogeneous* case, the support vector machine produces a vector $\mathbf{w} \in \mathbb{R}^n$, and classification of a point $\mathbf{x} \in \mathbb{R}^n$ is determined by $\text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle)$. In contrast, in the *non-homogeneous* case, the support vector machine produces a vector $\mathbf{w} \in \mathbb{R}^n$ and a value $b \in \bar{\mathbb{R}}$, and classification of a point $\mathbf{x} \in \mathbb{R}^n$ is determined by $\text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle + b)$. As is well-known, the latter case can easily be represented as a special case of the former, simply by the addition of one dimension, by fixing all data points to have a constant nonzero component in that extra dimension. However, the addition of a bias term can significantly affect the support vector machine algorithm and margin-based analysis thereof. Specifically, the notion of the *margin* of a point \mathbf{x} used in the definition of the support vector machine and analysis thereof is the *geometric margin*, $\frac{\langle \mathbf{w}, \mathbf{x} \rangle + b}{\|\mathbf{w}\|}$, corresponding to the Euclidean distance to the separating hyperplane. Since the bias term b is not included in the norm in the denominator, the definitions and results in the margin-based theory for non-homogeneous separators cannot quite be reduced to the homogeneous case by adding another dimension.

106 For this reason, throughout the presentation below, we will treat both the homogeneous and non-homogeneous cases in a unified fashion, by introducing a **global parameter** $c \in \{0, 1\}$. The case $c = 0$ will correspond to the homogeneous case, while $c = 1$ will correspond to the non-homogeneous case. In order to present both types of results simultaneously, we find it simplest to suppose

⁴ We should more properly be referring to \mathcal{D} as an ordered sequence of pairs (\mathbf{x}_i, y_i) , but have opted for this slight imprecision to retain the familiar phrase “dataset”.

111 the bias term b is always present, but that classification of a point $\mathbf{x} \in \mathbb{R}^n$ is
 112 determined by $\text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle + cb)$, so that the bias term b is simply ignored in the
 113 homogeneous case.

114 *2.1. Max-Margin Hyperplanes*

Definition 1 (Max-Margin Hyperplanes). For $m \in \mathbb{N}$, and a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i) \in \mathbb{R}^n \times \{-1, 1\} : i \in [m]\}$, we will write $(\hat{\mathbf{w}}, \hat{b}, \gamma) = \text{MMH}(\mathcal{D})$ to mean that $(\hat{\mathbf{w}}, \hat{b})$ represents the maximum-margin separator,

$$(\hat{\mathbf{w}}, \hat{b}) = \underset{\mathbf{w} \in \mathbb{S}^n, b \in \mathbb{R}}{\text{argmax}} \min_{i \in [m]} y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + cb),$$

115 and γ is the margin, $\gamma = \min_{i \in [m]} y_i (\langle \hat{\mathbf{w}}, \mathbf{x}_i \rangle + c\hat{b})$. If $c = 0$, for simplicity we
 116 define $\hat{b} = 0$, and in this case $\text{MMH}(\mathcal{D})$ is well-defined and unique as long as
 117 $\gamma > 0$. If $c = 1$, $\text{MMH}(\mathcal{D})$ is well-defined and unique as long as $\{y : (\mathbf{x}, y) \in$
 118 $\mathcal{D}\} = \{-1, 1\}$; for completeness, when this is not the case, we define $\hat{\mathbf{w}} \in \mathbb{S}^n$
 119 arbitrarily, $\hat{b} = y_1 \cdot \infty$, and $\gamma = \infty$.

120 Generally, if we wish to leave γ unspecified, we will simply write $(\hat{\mathbf{w}}, \hat{b}) =$
 121 $\text{MMH}(\mathcal{D})$. Alternatively, if we wish to leave $(\hat{\mathbf{w}}, \hat{b})$ unspecified, we will write
 122 $\gamma = \text{marg}(\mathcal{D})$.

123 Throughout this paper (with the exception of Section 8 on agnostic learning),
 124 we assume that (by definition of the term “dataset”) any dataset \mathcal{D} is *strictly*
 125 *linearly separable* — i.e., $\text{marg}(\mathcal{D}) > 0$. When $c = 1$, this is equivalent to linear
 126 separability, but for $c = 0$ it imposes an additional restriction.

127 **Definition 2** (Marginal Vectors). Suppose that \mathcal{D} is a dataset of m points, and
 128 $(\hat{\mathbf{w}}, \hat{b}, \gamma) = \text{MMH}(\mathcal{D})$. We say that $j \in [m]$ is a *marginal index* if $y_j (\langle \hat{\mathbf{w}}, \mathbf{x}_j \rangle +$
 129 $c\hat{b}) = \gamma$, and write $\mathcal{D}^{\text{marg}} \subseteq \mathcal{D}$ to denote the set of all $(\mathbf{x}_j, y_j) \in \mathcal{D}$ such that j
 130 is a marginal index; these are the *marginal vectors*. The *marginal vectors* are
 131 uniquely determined by \mathcal{D} — an immediate consequence of $(\hat{\mathbf{w}}, \hat{b})$ being uniquely
 132 defined.

133 **3. Main Results**

134 This section summarizes the main results of this work.

135 *3.1. Inductive SVM Error Bounds*

136 Fix a distribution over \mathbb{R}^n and a target $(\mathbf{w}^*, b^*) \in \mathbb{S}^n \times \bar{\mathbb{R}}$. The latter induces
 137 the target concept $f^* : \mathbb{R}^n \rightarrow \{-1, 1\}$ via $f^*(\mathbf{x}) = \text{sign}(\langle \mathbf{w}^*, \mathbf{x} \rangle + cb^*)$. Let
 138 $\mathbf{X}_1, \dots, \mathbf{X}_{m+1}$ be points be drawn i.i.d., labeled with $Y_i = f^*(\mathbf{X}_i)$, for $i \in [m+1]$.
 139 Denote by \mathcal{D}_{m+1} the full dataset consisting of the $m+1$ labeled points and by
 140 \mathcal{D}_m , this same dataset with the $(m+1)^{\text{th}}$ labeled example omitted. The *inductive*

141 SVM hypothesis \hat{h}_m predicts the label of \mathbf{X}_{m+1} based on \mathcal{D}_m using the max-
 142 margin hyperplane: $\hat{h}_m(\mathbf{x}; \mathcal{D}_m) = \text{sign}(\langle \hat{\mathbf{w}}, \mathbf{x} \rangle + c\hat{b})$, where $(\hat{\mathbf{w}}, \hat{b}) = \text{MMH}(\mathcal{D}_m)$.
 143 Associated with \hat{h}_m is its error

$$\text{err}(\hat{h}_m) = \mathbb{P}(\hat{h}_m(\mathbf{X}_{m+1}; \mathcal{D}_m) \neq Y_{m+1} \mid \mathbf{X}_1, \dots, \mathbf{X}_m) \quad (6)$$

144 and its expected error $\mathbb{E}[\text{err}(\hat{h}_m)]$, where the expectation is over the $\mathbf{X}_1, \dots, \mathbf{X}_m$.
 145 So that this classifier is uniquely defined, and a margin-based bound on its error
 146 rate is meaningful, we assume that the distribution of the \mathbf{X}_i samples is such
 147 that $\text{marg}(\mathcal{D}_{m+1}) > 0$ almost surely. This is true of *every* distribution when
 148 $c = 1$ (and hence is not really an assumption at all in that case), but it does
 149 impose a restriction on the distribution when $c = 0$.

150 We prove the following two results. The first provides a PAC generalization
 151 bound for the SVM, while the second bounds the expected error rate of the
 152 SVM.

153 **Theorem 3.** *Suppose that an iid sample $\mathcal{D} = \{(\mathbf{x}_i, y_i) \in \mathbb{R}^n \times \{-1, 1\} : i \in [m]\}$
 154 is contained in a ball of radius R and $(\hat{\mathbf{w}}, \hat{b}, \gamma) = \text{MMH}(\mathcal{D})$. Then, with proba-
 155 bility at least $1 - \delta$, we have*

$$\text{err}(\hat{h}) \leq \frac{2}{m} \left(5 \left[\frac{R}{\gamma} \right]^2 \log_2 \frac{em\gamma^2}{R^2} + \log_2 \left(\frac{\pi^4}{9\delta} \left[\frac{R}{\gamma} \right]^2 \right) \right),$$

156 where $\hat{h} : \mathbb{R}^n \rightarrow \{-1, 1\}$ is defined by $\hat{h}(\mathbf{x}) = \text{sign}(\langle \hat{\mathbf{w}}, \mathbf{x} \rangle + \hat{b})$.

157 *Remark:* To our knowledge, the bounds appearing in published literature have
 158 a $\log^2(m)/m$ dependence on sample size, from which the above result shaves off
 159 a logarithmic factor. John Shawe-Taylor [7] informs us that such a result follows
 160 from Zhang's bounds on covering numbers for linear function classes [10], but
 161 the argument we will give is considerably more elementary and yields better
 162 constants.

Theorem 4. *For $\mathcal{D}_m, \mathcal{D}_{m+1}$, and \hat{h}_m as defined above, let $\gamma_{m+1} = \text{marg}(\mathcal{D}_{m+1})$,
 and let $r_{m+1} = \max_{i \in [m+1]} \|\mathbf{X}_i\|$. Then*

$$\mathbb{P}(\hat{h}_m(\mathbf{X}_{m+1}; \mathcal{D}_m) \neq Y_{m+1} \mid \gamma_{m+1}, r_{m+1}) \leq \frac{1}{m+1} \min \left\{ n + c, \frac{(2+6c)r_{m+1}^2}{\gamma_{m+1}^2} \right\}, \quad (7)$$

$$\mathbb{E}[\text{err}(\hat{h}_m)] \leq \frac{1}{m+1} \mathbb{E} \left[\min \left\{ n + c, \frac{(2+6c)r_{m+1}^2}{\gamma_{m+1}^2} \right\} \right]. \quad (8)$$

163 As discussed above, in the case $c = 1$, this result was first established by
 164 [1], via a different argument (see below for discussion of the differences). To
 165 our knowledge, this is the first publication establishing this result for the case
 166 $c = 0$, which (as discussed above) is in many respects quite a different setting.
 167 Our proof is able to handle both cases simultaneously. Our new proof of this
 168 result is presented in Section 6 below.

169 *Deficiencies in Reducing the Non-homogeneous Case to the Homogeneous Case.*
170 As mentioned above, the margin analysis that one finds in most standard treat-
171 ments only addresses the *homogeneous case*, reasoning that one can always
172 reduce the non-homogeneous case to the homogeneous case simply by adding
173 a dimension and fixing that coordinate to 1 in all the data points. With the
174 above bounds in hand, we can now discuss quantitatively why that approach
175 sometimes leads to significantly larger bounds on the error rate. Specifically,
176 first consider a data set \mathcal{D}'_m of points (\mathbf{x}'_i, y_i) with $\|\mathbf{x}'_i\| = 1$, such that for
177 $(\mathbf{w}', b', \gamma') = \text{MMH}(\mathcal{D}'_m)$, we have $b' = 0$. Now Theorem 3 would supply a
178 bound $O\left(\frac{1}{m} \left(\frac{1}{(\gamma')^2} \log(m(\gamma')^2) + \log \frac{1}{\delta(\gamma')^2}\right)\right)$, regardless of whether we treat
179 this as the homogeneous or non-homogeneous solution (as both have zero bias).
180 However, if we were to uniformly shift this dataset, without changing the ge-
181 ometric margin of the SVM solution, we suddenly find a dramatic difference
182 between the direct analysis of the non-homogeneous case in Theorem 3 and the
183 naïve reduction technique implicit in the traditional analysis. Specifically, let
184 \mathcal{D}_m be the dataset of points (\mathbf{x}_i, y_i) , where $\mathbf{x}_i = \mathbf{x}'_i + (R - 1)\mathbf{w}'$, where R is
185 a large positive value. Then letting $(\hat{\mathbf{w}}, \hat{b}, \gamma) = \text{MMH}(\mathcal{D}_m)$, we may note that
186 $(\hat{\mathbf{w}}, \hat{b}, \gamma) = (\mathbf{w}', 1 - R, \gamma')$. Thus, since the geometric margin is unchanged,
187 and the samples are contained in a ball of radius R , Theorem 3 provides a
188 bound $O\left(\frac{1}{m} \left(\frac{R^2}{\gamma^2} \log \frac{m\gamma^2}{R^2} + \log \frac{R^2}{\delta\gamma^2}\right)\right)$. However, if we were instead to add a
189 dimension, with coordinate value fixed to 1, and treat this scenario in the *ho-*
190 *mogeneous* case, then the maximum margin separator would have weight vec-
191 tor $\frac{(\hat{\mathbf{w}}, 1 - R)}{\|(\hat{\mathbf{w}}, 1 - R)\|}$, and would have margin $\min_i \left| \frac{1}{\|(\hat{\mathbf{w}}, 1 - R)\|} \langle (\hat{\mathbf{w}}, 1 - R), (\mathbf{x}_i, 1) \rangle \right| =$
192 $\frac{\gamma}{\sqrt{1 + (1 - R)^2}}$. Thus, with this naïve reduction-to-homogeneous approach, for
193 large R , the bound one obtains by plugging into a result such as Theorem 3
194 is $O\left(\frac{1}{m} \left(\frac{R^4}{\gamma^2} \log \frac{m\gamma^2}{R^4} + \log \frac{R^4}{\delta\gamma^2}\right)\right)$, which is larger than the result above directly
195 analyzing the non-homogeneous case by roughly a factor of R^2 . Thus, we see
196 that it can be extremely important to explicitly treat the bias term separately
197 when bounding the error rate. Furthermore, as mentioned above, the value of
198 the margin in the above bounds corresponds to the same value appearing in the
199 objective function of the support vector machine (i.e., the geometric margin),
200 while this is not the case if we treat the bias term as part of the weight vector (as
201 in the reduction-to-homogeneous approach). Thus, in addition to sometimes be-
202 ing quantitatively tighter, the bounds above obtained by treating the bias term
203 separately directly motivate the support vector machine optimization problem.

204 3.2. Transductive SVM Error Bounds

205 Our strategy for the transductive error bound shares several features with
206 the inductive case. We begin with the same setting as in Section 6.2: a target
207 $(\mathbf{w}^*, b^*) \in \mathbb{S}^n \times \mathbb{R}$ with its induced target concept $f^* : \mathbb{R}^n \rightarrow \{-1, 1\}$, and the
208 i.i.d. dataset $\mathcal{D}_{m+1} = \{(\mathbf{X}_i, Y_i) : i \in [m + 1]\}$, as well as its “abridged” version
209 \mathcal{D}_m . We also continue the assumption that, in the case $c = 0$, the distribution
210 of \mathbf{X}_i is such that $\text{marg}(\mathcal{D}_{m+1}) > 0$ almost surely.

The *transductive SVM hypothesis* \hat{h}_m predicts the label of \mathbf{X}_{m+1} based on \mathcal{D}_m as follows:

$$\hat{h}_m(\mathbf{x}; \mathcal{D}_m) = \operatorname{argmax}_{y \in \{-1, 1\}} \sup_{\mathbf{w} \in \mathbb{S}^n, b \in \mathbb{R}} \min \left\{ y (\langle \mathbf{w}, \mathbf{x} \rangle + cb), \min_{i \in [m]} Y_i (\langle \mathbf{w}, \mathbf{X}_i \rangle + cb) \right\}.$$

211 The error rate, $\operatorname{err}(\hat{h}_m)$, of this classifier is defined as above in (6).

212 We establish the following result bounding the expected error rate of the
213 transductive SVM.

Theorem 5. For $\mathcal{D}_m, \mathcal{D}_{m+1}, \hat{h}_m$ as defined above, letting $\gamma_{m+1} = \operatorname{marg}(\mathcal{D}_{m+1})$, and $r_{m+1} = \max_{i \in [m+1]} \|\mathbf{X}_i\|$, we have

$$\mathbb{P} \left(\hat{h}_m(\mathbf{X}_{m+1}; \mathcal{D}_m) \neq Y_{m+1} \mid \gamma_{m+1}, r_{m+1} \right) \leq \frac{1}{m+1} \min \left\{ n + c, \frac{(1+3c)r_{m+1}^2}{\gamma_{m+1}^2} \right\}, \quad (9)$$

$$\mathbb{E} \left[\operatorname{err}(\hat{h}_m) \right] \leq \frac{1}{m+1} \mathbb{E} \left[\min \left\{ n + c, \frac{(1+3c)r_{m+1}^2}{\gamma_{m+1}^2} \right\} \right]. \quad (10)$$

214 The proof of this result follows a similar outline as the analysis of inductive
215 SVM, and is presented in Section 7.

216 In particular, recalling the lower bound (1), which applies to learning margin-
217 γ homogeneous linear separators, the bound in Theorem 5 implies that in the
218 homogeneous case, the transductive SVM is asymptotically minimax optimal.

219 4. SVM PAC Generalization Bound

220 The main result of this section is a proof of Theorem 3. To facilitate the
221 proof, we define the following parametrized family of concepts. For $R, \Lambda > 0$,
222 consider all $h : \mathbb{R}^n \times \{-1, 1\} \rightarrow \{-1, 1\}$ of the form

$$(\mathbf{x}, y) \mapsto \begin{cases} \operatorname{sign}(y(\langle \mathbf{w}, \mathbf{x} \rangle + b)), & \|\mathbf{x}\| \leq R, |\langle \mathbf{w}, \mathbf{x} \rangle + b| \geq 1 \\ -1, & \text{else,} \end{cases}$$

223 where $(\mathbf{w}, b) \in \mathbb{R}^n \times \mathbb{R}$ range over all $\|\mathbf{w}\| \leq \Lambda$ (and b is arbitrary).

224 A hypothesis $h \in \mathcal{C}_{R, \Lambda}$ is said to be *consistent* with a labeled sample $\mathcal{D}_m =$
225 $\{(\mathbf{X}_i, Y_i) : i \in [m]\}$ if $h(\mathbf{X}_i, Y_i) = 1$ for all $i \in [m]$. These are essentially the
226 *gap-tolerant classifiers* [11].

227 **Lemma 6.** The VC-dimension of $\mathcal{C}_{R, \Lambda}$ is at most $(2R\Lambda + 1)^2$.

228 **Remark:** To establish this lemma, we will closely follow the proof of [4, Theo-
229 rem 4.2]. The differences are that the latter (i) does not allow a bias term b , (ii)
230 defines a sample-dependent concept class, which precludes invoking standard
231 PAC bounds (which require that the concept class be fixed in advance of seeing
232 the sample) and (iii) has a concept class defined over the points \mathbf{x} as opposed
233 to pairs (\mathbf{x}, y) .

234 *Proof of Lemma 6.* Suppose that $\mathcal{C}_{R,\Lambda}$ shatters some set $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_d, y_d)\}$
 235 of pairs, for some $d \in \mathbb{N}$. This implies, in particular, that $\|\mathbf{x}_i\| \leq R$, $i \in [d]$. It
 236 also implies that, for all $\mathbf{s} \in \{-1, 1\}^d$, there is a $(\mathbf{w}, b) \in \mathbb{R}^n \times \mathbb{R}$, $\|\mathbf{w}\| \leq \Lambda$, such
 237 that

$$1 \leq s_i(y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b)) = s_i(\langle \mathbf{w}, y_i \mathbf{x}_i \rangle + by_i), \quad i \in [d].$$

Note that, for any (\mathbf{w}, b) satisfying this inequality with $\|\mathbf{w}\| \leq \Lambda$, if $b > R\Lambda + 1$,
 then $1 \leq s_i(y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + R\Lambda + 1))$ as well, and if $b < -(R\Lambda + 1)$, then $1 \leq$
 $s_i(y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle - (R\Lambda + 1)))$ as well. Thus, without loss of generality, we may
 suppose $|b| \leq R\Lambda + 1$. Summing up the inequalities over $i \in [d]$, we have

$$\begin{aligned} d &\leq \sum_{i \in [d]} s_i(\langle \mathbf{w}, y_i \mathbf{x}_i \rangle + by_i) = \left\langle \mathbf{w}, \sum_{i \in [d]} s_i y_i \mathbf{x}_i \right\rangle + b \sum_{i \in [d]} s_i y_i \\ &\leq \Lambda \left\| \sum_{i \in [d]} s_i y_i \mathbf{x}_i \right\| + (R\Lambda + 1) \left| \sum_{i \in [d]} s_i y_i \right|. \end{aligned}$$

Letting \mathbf{s} be uniformly drawn from $\{-1, 1\}^d$ and taking expectations (noting
 that the $\{s_i\}$ are independent and $\mathbb{E}[s_i] = 0$), we have

$$\begin{aligned} d &\leq \Lambda \mathbb{E} \left\| \sum_i s_i y_i \mathbf{x}_i \right\| + (R\Lambda + 1) \mathbb{E} \left| \sum_i s_i y_i \right| \\ &\leq \Lambda \sqrt{\mathbb{E} \left\| \sum_i s_i y_i \mathbf{x}_i \right\|^2} + (R\Lambda + 1) \sqrt{\mathbb{E} \left(\sum_i s_i y_i \right)^2} \\ &= \Lambda \sqrt{\sum_i \|y_i \mathbf{x}_i\|^2} + (R\Lambda + 1) \sqrt{\sum_i y_i^2} \\ &\leq \Lambda R \sqrt{d} + (R\Lambda + 1) \sqrt{d} = (2R\Lambda + 1) \sqrt{d}. \end{aligned}$$

238 Solving, $d \leq (2R\Lambda + 1)^2$. □

239 *Proof of Theorem 3.* Recall a classic VC-based generalization bound for consist-
 240 ent binary classifiers [12]: with probability at least $1 - \delta$, a classifier consistent
 241 with a sample of size m chosen from a concept class with VC-dimension d
 242 achieves generalization error at most $\frac{2}{m} (d \log_2 \frac{2em}{d} + \log_2 \frac{2}{\delta})$.

243 Our learner's task is to match the function $f : \mathbb{R}^n \times \{-1, 1\} \rightarrow \{-1, 1\}$ given
 244 by $f(\mathbf{x}, y) = 1$ on the labeled sample using concepts $h \in \mathcal{C}_{R,\Lambda}$. We are going
 245 to invoke a standard (double) stratification argument [13] over $\|\mathbf{w}\|$ and $\|\mathbf{x}\|$.
 246 Define the lattice of concept classes $\mathcal{C}_{i,j}$, where $\mathcal{C}_{i,j} \subseteq \mathcal{C}_{i',j'}$ whenever $i \leq i'$
 247 and $j \leq j'$. This lattice is defined *in advance* of seeing any sample. Now consider a
 248 learner who receives a training sample $S \subset \mathbb{R}^n$, contained in some ball of radius
 249 R . There exists a consistent hyperplane with margin γ iff there is a consistent
 250 $h \in \mathcal{C}_{\lceil R \rceil, \lceil 1/\gamma \rceil}$. Define $p_i = 6/(\pi i)^2$, for $i = 1, 2, \dots$, and q_j analogously.

251 Then, by the stratification argument, with probability at least $1 - \delta$, any
 252 γ -margin hyperplane consistent with a sample of size m contained in radius R
 253 achieves generalization error at most

$$\frac{2}{m} \left(\left(2 \lceil R \rceil \left\lceil \frac{1}{\gamma} \right\rceil + 1 \right)^2 \log_2 \frac{2em}{(2\lceil R \rceil \lceil 1/\gamma \rceil + 1)^2} + \log_2 \frac{2}{\delta q_{\lceil R \rceil p_{\lceil 1/\gamma \rceil}}}, \right),$$

254 from which the stated bound follows immediately, using $2uv \leq 2 \lceil u \rceil \lceil v \rceil + 1 \leq$
 255 $5 \lceil uv \rceil$, which is valid for all $u, v \geq 0$. \square

256 5. Representation by Lagrange multipliers

257 The following lemma summarizes some well-known facts about max-margin
 258 hyperplanes and their induced support vectors. They may be seen as conse-
 259 quences of the strong duality and complementary slackness [14, 15], see also
 260 [16, 4]. The representation of $\text{MMH}(\mathcal{D})$ in terms of Lagrange multipliers α ,
 261 and properties thereof, as described in this lemma, will be vital to our analysis
 262 below.

263 **Lemma 7.** *For any $m \in \mathbb{N}$, consider a data set $\mathcal{D} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\}$,
 264 and put $\mathbf{z}_i := y_i \mathbf{x}_i \in \mathbb{R}^n, i \in [m]$. Suppose that $(\hat{\mathbf{w}}, \hat{b}, \gamma) = \text{MMH}(\mathcal{D})$ with
 265 $0 < \gamma < \infty$. Then:*

- 266 (i) *There exists an $\alpha \in \mathbb{R}_+^m$ such that $\hat{\mathbf{w}} = \sum_{i=1}^m \alpha_i \mathbf{z}_i$
 267 (meaning: the normal vector $\hat{\mathbf{w}}$ of the max-margin hyperplane lies in the
 268 conical hull of the data vectors).*
- 269 (ii) *The α in (i) may be chosen to satisfy $\alpha_i \neq 0 \implies \langle \hat{\mathbf{w}}, \mathbf{z}_i \rangle + y_i \hat{b} = \gamma, i \in [m]$
 270 (meaning: the margin is achieved at the support vectors).*
- 271 (iii) *The α in (i,ii) may be chosen to satisfy $c \sum_{i=1}^m \alpha_i y_i = 0$
 272 (meaning: in the non-homogeneous case, the sums of α multipliers for
 273 positive and negative examples are equal).*
- 274 (iv) *For any α as in (i,ii,iii), putting $\mathcal{D}^{\text{supp}} = \{(\mathbf{x}_i, y_i) : i \in \text{supp}(\alpha)\}$, we have
 275 $(\hat{\mathbf{w}}, \hat{b}, \gamma) = \text{MMH}(\mathcal{D}^{\text{supp}})$
 276 (meaning: the non-support vectors may be omitted from the data set with-
 277 out affecting the max-margin hyperplane).*
- (v) *Assuming $m > 1$, choose $i \in [m]$ and let $(\hat{\mathbf{w}}_{-i}, \hat{b}_{-i}, \gamma_{-i}) = \text{MMH}(\mathcal{D}_{-i})$.
 Then
 278 $\gamma_{-i} > \gamma \iff (\hat{\mathbf{w}}_{-i}, \hat{b}_{-i}) \neq (\hat{\mathbf{w}}, \hat{b}) \implies y_i (\langle \hat{\mathbf{w}}, \mathbf{x}_i \rangle + \hat{b}) = \gamma \iff (\mathbf{x}_i, y_i) \in \mathcal{D}^{\text{marg}}$
 279 (meaning: omitting the i^{th} example increases the margin iff it changes the
 280 optimal hyperplane, and this implies that the omitted point was a marginal
 vector).*

281 Given any $\alpha \in \mathbb{R}^m$ as described in Lemma 7(i, iv), we generally denote
 282 $\mathcal{D}^{\text{supp}} = \{(\mathbf{x}_i, y_i) : i \in \text{supp}(\alpha)\}$, the set of *support vectors* (with respect to α).
 283 Note, however, that as the vector α in Lemma 7 is not guaranteed to be unique,
 284 or even to have unique support, the set $\mathcal{D}^{\text{supp}}$ is generally not uniquely defined.

285 6. Inductive SVM Expected Error Bound

286 Before getting into the details of our proof, we first briefly discuss some
287 important similarities and differences in our approach compared to the previous
288 proof of [1]. The proof of [1] studies the *leave-one-out* cross validation error of
289 the algorithm, which is known to be an unbiased estimator of the error rate.
290 They bound this value in terms of the number of “essential” support vectors
291 (whose inclusion is required by any solution to the SVM optimization problem),
292 and then bound this number by $1/\gamma_{m+1}^2$. The proof of this latter bound lower-
293 bounds the Lagrange multiplier for any data point counted as a mistake in the
294 leave-one-out estimator. It does so by considering the effect on the Lagrange
295 multipliers of the other points induced by fixing that point’s multiplier to 0
296 in the SVM dual optimization problem, and analyzing the effect on the dual
297 objective function.

298 Like the original proof of [1], our new proof also examines the leave-one-out
299 cross validation error of the SVM, and relates this to the number of essential
300 support vectors, which we then bound by a value $\propto 1/\gamma_{m+1}^2$ by lower-bounding
301 the Lagrange multipliers of points counted as a mistake in the leave-one-out
302 estimator. However, our proof diverges from that of [1] in this last step. Specif-
303 ically, rather than analyzing the Lagrange multipliers of the solution to the SVM
304 dual optimization problem with the point held out, we are able to lower-bound
305 the Lagrange multipliers of mistake points by analyzing the effect of leaving out
306 that point, in terms of the weight vector in the solution to the *primal* optimiza-
307 tion problem. This yields new insights into the behavior of the primal solutions
308 in support vector machines, which may themselves be of interest.

309 Our approach is also quite flexible, and in particular allows us to simultane-
310 ously analyze the homogeneous (zero bias term) and non-homogeneous variants
311 of SVM, yielding smaller constant factors in the former case, which was not
312 covered by the original proof of [1].

313 As we will see in the next section, the approach also easily extends to the
314 analysis of *transductive* SVM, where we also obtain bounds on the expected
315 error rate, for both the homogeneous and non-homogeneous cases, which match
316 the lower bound (1) up to constant factors. To our knowledge, this is the first
317 publication of a proof that transductive SVM obtains the minimax rate. We
318 note, however, that one can modify the argument of [1] to obtain a similar result
319 for transductive SVM, though again only for the non-homogeneous case.

320 6.1. Bounding the Number of Leave-one-out Mistake Vectors

321 As mentioned, our basic strategy toward bounding the expected error rate
322 of the SVM is to analyze its leave-one-out cross validation error rate, which
323 (when the test point is included in the data set) is known to be an unbiased
324 estimator of the expected error rate. Toward this end, we now define the set of
325 *leave-one-out mistake vectors* — corresponding to the data points on which a

326 mistake is made when they are held out.⁵

327 **Definition 8** (Leave-one-out Mistake Vectors). *Given a dataset \mathcal{D} as above, we*
 328 *say that $\ell \in [m]$ is a leave-one-out mistake index if $(\hat{\mathbf{w}}_{-\ell}, \hat{b}_{-\ell}) = \text{MMH}(\mathcal{D}_{-\ell})$*
 329 *satisfies $y_\ell (\langle \hat{\mathbf{w}}_{-\ell}, \mathbf{x}_\ell \rangle + c\hat{b}_{-\ell}) \leq 0$. In other words, upon removing \mathbf{x}_ℓ from \mathcal{D} ,*
 330 *the resulting maximum margin separator misclassifies \mathbf{x}_ℓ (or possibly has \mathbf{x}_ℓ on*
 331 *the separator). Let $\mathcal{D}^{\text{LOOM}} \subseteq \mathcal{D}$ denote the set of all $(\mathbf{x}_\ell, y_\ell) \in \mathcal{D}$ such that*
 332 *$\ell \in [m]$ is a leave-one-out mistake index; these are the leave-one-out mistake*
 333 *vectors.*

334 The rest of this section is devoted to proving the following theorem, which
 335 bounds the number of leave-one-out mistake vectors in terms of the dimension
 336 n and the margin γ .

337 **Theorem 9.** *Fix any $m \in \mathbb{N}$ with $m \geq 2$, and any $\mathcal{D} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\}$,*
 338 *and let $(\hat{\mathbf{w}}, \hat{b}, \gamma) = \text{MMH}(\mathcal{D})$. Let $r \in \mathbb{R}_+$ be such that $\max_{i \in [m]} \|\mathbf{x}_i\| \leq r$.*
 339 *Assuming $\gamma > 0$, we have $\text{card}(\mathcal{D}^{\text{LOOM}}) \leq \min \left\{ n + c, \frac{(2+6c)r^2}{\gamma^2} \right\}$.*

340 The proof of this theorem relies on the following lemma, which lower-bounds
 341 the Lagrange multipliers α from Lemma 7 associated with vectors whose margin
 342 can be reduced without reducing the margin on the remaining points.

343 **Lemma 10.** *Let \mathcal{D} , $\{\mathbf{z}_i\}$, $(\hat{\mathbf{w}}, \hat{b})$, γ , and α be as in Lemma 7(i,ii,iii), and*
 344 *suppose $0 < \gamma < \infty$. Let $r \in (0, \infty)$ be such that $\max_{i \in [m]} \|\mathbf{x}_i\| \leq r$. Fix*
 345 *any $\epsilon \in (-\infty, \gamma)$, and if $c = 1$, then also suppose $\epsilon \geq \gamma - 4r^2/\gamma$. For any*
 346 *$d \in [m]$, if there exists $(\mathbf{w}_d, b_d) \in \mathbb{S}^n \times \mathbb{R}$ such that $y_d (\langle \mathbf{w}_d, \mathbf{x}_d \rangle + cb_d) \leq \epsilon$ and*
 347 *$\min_{j \in [m] \setminus \{d\}} y_j (\langle \mathbf{w}_d, \mathbf{x}_j \rangle + cb_d) \geq \gamma$, then $\alpha_d \geq \frac{1}{(2+6c)r^2} (\gamma - \epsilon)$.*

Proof. Let $\gamma_d = y_d (\langle \mathbf{w}_d, \mathbf{x}_d \rangle + cb_d)$, and note that $\gamma_d \leq \epsilon < \gamma$. Lemma 7(i,iii) and Lemma 18 imply that

$$\begin{aligned} \langle \hat{\mathbf{w}}, \mathbf{w}_d \rangle &= \sum_{j=1}^m \alpha_j \langle \mathbf{w}_d, \mathbf{z}_j \rangle = \alpha_d \langle \mathbf{w}_d, \mathbf{z}_d \rangle + \sum_{j \in [m] \setminus \{d\}} \alpha_j \langle \mathbf{w}_d, \mathbf{z}_j \rangle \\ &= \alpha_d (\gamma_d - y_d c b_d) + \sum_{j \in [m] \setminus \{d\}} \alpha_j \langle \mathbf{w}_d, \mathbf{z}_j \rangle \geq \alpha_d (\gamma_d - y_d c b_d) + \sum_{j \in [m] \setminus \{d\}} \alpha_j (\gamma - y_j c b_d) \\ &= \alpha_d (\gamma_d - \gamma) + \gamma \sum_{j=1}^m \alpha_j - b_d c \sum_{j=1}^m \alpha_j y_j = \alpha_d (\gamma_d - \gamma) + 1. \end{aligned}$$

If it is also true that $\langle \hat{\mathbf{w}}, \mathbf{w}_d \rangle \leq 1 - \frac{1}{(2+6c)r^2} (\gamma - \gamma_d)^2$, then altogether we have

$$\alpha_d \geq \frac{1}{(2+6c)r^2} (\gamma - \gamma_d) \geq \frac{1}{(2+6c)r^2} (\gamma - \epsilon).$$

⁵For simplicity, we also include data points (\mathbf{x}_i, y_i) which, when held out, are *borderline* predictions (i.e., those on the $\text{MMH}(\mathcal{D}_{-i})$ separator). Since our purpose below is to upper bound the number of leave-one-out mistakes, this relaxation is benign.

348 Otherwise, if $\langle \hat{\mathbf{w}}, \mathbf{w}_d \rangle > 1 - \frac{1}{(2+6c)r^2}(\gamma - \gamma_d)^2$, then supposing $\|\mathbf{x}_d\| > 0$,
 349 $1 - \frac{1}{(2+6c)r^2}(\gamma - \gamma_d)^2 \geq 1 - \frac{1}{(2+6c)\|\mathbf{x}_d\|^2}(\gamma - \gamma_d)^2$, so that Lemma 16 from Ap-
 350 pendix Appendix A implies

$$\langle \hat{\mathbf{w}}, \mathbf{z}_d \rangle - \langle \mathbf{w}_d, \mathbf{z}_d \rangle < \frac{1}{1+c}(\gamma - \gamma_d).$$

351 This inequality is also trivially satisfied if $\|\mathbf{x}_d\| = 0$. But since $\langle \hat{\mathbf{w}}, \mathbf{z}_d \rangle + y_d \hat{c}b \geq \gamma$
 352 and $\langle \mathbf{w}_d, \mathbf{z}_d \rangle + y_d cb_d = \gamma_d$, this implies

$$\begin{aligned} y_d c(\hat{b} - b_d) &> (\langle \hat{\mathbf{w}}, \mathbf{z}_d \rangle + y_d \hat{c}b) - (\langle \mathbf{w}_d, \mathbf{z}_d \rangle + y_d cb_d) - \frac{1}{1+c}(\gamma - \gamma_d) \\ &\geq (\gamma - \gamma_d) - \frac{1}{1+c}(\gamma - \gamma_d) = \frac{c}{1+c}(\gamma - \gamma_d). \end{aligned}$$

In particular, since this can only occur with $c = 1$, this completes the proof for the case $c = 0$. Now for the case $c = 1$, suppose $j \in [m] \setminus \{d\}$ is such that $y_j = y_d$. If $\|\mathbf{x}_j\| > 0$, then since $\langle \hat{\mathbf{w}}, \mathbf{w}_d \rangle > 1 - \frac{1}{8r^2}(\gamma - \gamma_d)^2 \geq 1 - \frac{1}{8\|\mathbf{x}_j\|^2}(\gamma - \gamma_d)^2$, Lemma 16 from Appendix Appendix A implies $\langle \mathbf{w}_d, \mathbf{z}_j \rangle - \langle \hat{\mathbf{w}}, \mathbf{z}_j \rangle < \frac{1}{2}(\gamma - \gamma_d)$. This inequality is also trivially satisfied if $\|\mathbf{x}_j\| = 0$. Thus, we have

$$\begin{aligned} \langle \hat{\mathbf{w}}, \mathbf{z}_j \rangle + y_j \hat{b} &= \langle \hat{\mathbf{w}}, \mathbf{z}_j \rangle + y_d \hat{b} > \left(\langle \mathbf{w}_d, \mathbf{z}_j \rangle - \frac{1}{2}(\gamma - \gamma_d) \right) + \left(y_d b_d + \frac{1}{2}(\gamma - \gamma_d) \right) \\ &= \langle \mathbf{w}_d, \mathbf{z}_j \rangle + y_j b_d \geq \gamma. \end{aligned}$$

353 In particular, this implies $(\mathbf{x}_j, y_j) \notin \mathcal{D}^{\text{marg}}$. Together with Lemma 7(ii), this
 354 implies $(\mathbf{x}_j, y_j) \notin \mathcal{D}^{\text{supp}}$ (defined with respect to $\boldsymbol{\alpha}$). Thus, if $\langle \hat{\mathbf{w}}, \mathbf{w}_d \rangle > 1 -$
 355 $\frac{1}{8r^2}(\gamma - \gamma_d)^2$, then every $j \in [m] \setminus \{d\}$ with $\alpha_j > 0$ has $y_j \neq y_d$. But together
 356 with Lemma 7(iii) and Lemma 18, this implies

$$\alpha_d = \sum_{j \in [m] \setminus \{d\}} \alpha_j = -\alpha_d + \sum_{j=1}^m \alpha_j = \frac{1}{\gamma} - \alpha_d,$$

357 so that $\alpha_d = \frac{1}{2\gamma} \geq \frac{1}{8r^2}(\gamma - \epsilon)$. \square

358 In particular, this straightforwardly implies the following corollary, lower-
 359 bounding the α_ℓ values for leave-one-out mistake indices ℓ .

360 **Corollary 11.** *Let \mathcal{D} , $\{\mathbf{z}_i\}$, $(\hat{\mathbf{w}}, \hat{b})$, γ , $\boldsymbol{\alpha}$, and r be as in Lemma 10. Then*
 361 $(\mathbf{x}_\ell, y_\ell) \in \mathcal{D}^{\text{LOOM}} \implies \alpha_\ell \geq \frac{1}{(2+6c)r^2}\gamma$, $\ell \in [m]$.

362 *Proof.* The claim is vacuously true if $\mathcal{D}^{\text{LOOM}} = \emptyset$ or $\gamma = 0$ (since $\boldsymbol{\alpha} \in \mathbb{R}_+^m$), so
 363 suppose that $\mathcal{D}^{\text{LOOM}}$ contains some $(\mathbf{x}_\ell, y_\ell)$, and that $\gamma > 0$. Next, note that
 364 the fact that $\gamma < \infty$ implies that there exist $j, j' \in [m]$ with $y_j \neq y_{j'}$. In
 365 particular, this means $\exists \tau \in [0, 1]$ such that, denoting $\mathbf{x}_\tau = \tau \mathbf{x}_j + (1 - \tau) \mathbf{x}_{j'}$,
 366 $\langle \hat{\mathbf{w}}, \mathbf{x}_\tau \rangle + \hat{c}b = 0$. Thus, since $\mathbf{x} \mapsto |\langle \hat{\mathbf{w}}, \mathbf{x} \rangle + \hat{c}b|$ is the Euclidean distance from
 367 \mathbf{x} to the *closest* point \mathbf{x}_0 with $\langle \hat{\mathbf{w}}, \mathbf{x}_0 \rangle + \hat{c}b = 0$, a triangle inequality implies

368 $|\langle \hat{\mathbf{w}}, \mathbf{x}_j \rangle + \hat{c}b| + |\langle \hat{\mathbf{w}}, \mathbf{x}_{j'} \rangle + \hat{c}b| \leq \|\mathbf{x}_j - \mathbf{x}_\tau\| + \|\mathbf{x}_{j'} - \mathbf{x}_\tau\| = \|\mathbf{x}_j - \mathbf{x}_{j'}\| \leq$
 369 $\|\mathbf{x}_j\| + \|\mathbf{x}_{j'}\| \leq 2r$. Since $|\langle \hat{\mathbf{w}}, \mathbf{x}_j \rangle + \hat{c}b| + |\langle \hat{\mathbf{w}}, \mathbf{x}_{j'} \rangle + \hat{c}b| \geq 2\gamma$, this implies $\gamma \leq r$.

Let $(\hat{\mathbf{w}}_{-\ell}, \hat{b}_{-\ell}) = \text{MMH}(\mathcal{D}_{-\ell})$, and note (from Definition 8) that

$$y_\ell (\langle \hat{\mathbf{w}}_{-\ell}, \mathbf{x}_\ell \rangle + \hat{c}b_{-\ell}) \leq 0,$$

and, since removing a point cannot decrease the maximum achievable margin,

$$\min_{j \in [m] \setminus \{\ell\}} y_j (\langle \hat{\mathbf{w}}_{-\ell}, \mathbf{x}_j \rangle + \hat{c}b_{-\ell}) \geq \gamma.$$

370 Thus, since $0 < \gamma \leq r$ implies $0 \in [\gamma - 4r^2/\gamma, \gamma)$, the result follows from
 371 Lemma 10 (taking $\epsilon = 0$). \square

372 We are now ready for the proof of Theorem 9.

373 *Proof of Theorem 9.* If $\gamma = \infty$ (which can only happen if $c = 1$), then it must
 374 be that every $(\mathbf{x}_i, y_i) \in \mathcal{D}$ has the same y_i . Since $m \geq 2$, this implies that
 375 every $i \in [m]$ has $(\hat{\mathbf{w}}_{-i}, \hat{b}_{-i}) = \text{MMH}(\mathcal{D}_{-i})$ with $\hat{b}_{-i} = y_1 \infty = y_i \infty$, so that
 376 $y_i (\langle \hat{\mathbf{w}}_{-i}, \mathbf{x}_i \rangle + \hat{c}b_{-i}) = \infty > 0$, and hence $(\mathbf{x}_i, y_i) \notin \mathcal{D}^{\text{LOOM}}$: that is, $\mathcal{D}^{\text{LOOM}} = \emptyset$.
 377 The result trivially follows in this case. Furthermore, note that if $r = 0$ (which
 378 again can only happen if $c = 1$, due to the $\gamma > 0$ assumption), then every
 379 $(\mathbf{x}_i, y_i) \in \mathcal{D}$ has the same \mathbf{x}_i . Together with the linear separability assumption,
 380 this again implies that every $(\mathbf{x}_i, y_i) \in \mathcal{D}$ has the same y_i , so that $\gamma = \infty$, and
 381 hence, as just established, the result trivially holds in this case.

For the remaining case, suppose $0 < \gamma < \infty$ and $r > 0$, and put $k = \text{card}(\mathcal{D}^{\text{LOOM}})$. The claim is trivial if $k = 0$, so assume $k \geq 1$ and let $\{i_1, \dots, i_k\} \subseteq [m]$ be the leave-one-out mistake indices:

$$\mathcal{D}^{\text{LOOM}} = \{(\mathbf{x}_{i_1}, y_{i_1}), \dots, (\mathbf{x}_{i_k}, y_{i_k})\}.$$

Put $\mathbf{z}_i = y_i \mathbf{x}_i$, $i \in [m]$. Lemma 17 implies the existence of $\boldsymbol{\alpha} \in \mathbb{R}_+^m$ satisfying the conditions (i,ii,iii) of Lemma 7, such that the vectors $\{(\mathbf{x}_i, c) : i \in \text{supp}(\boldsymbol{\alpha})\}$ are linearly independent. Furthermore, Corollary 11 implies that for any such $\boldsymbol{\alpha}$, $\alpha_{i_j} \geq \frac{1}{(2+6c)r^2} \gamma > 0$, $j \in [k]$. Thus, any leave-one-out mistake index i_j must be in $\text{supp}(\boldsymbol{\alpha})$, and hence

$$\{(\mathbf{x}_i, c) : (\mathbf{x}_i, y_i) \in \mathcal{D}^{\text{LOOM}}\} \subseteq \{(\mathbf{x}_i, c) : i \in \text{supp}(\boldsymbol{\alpha})\}.$$

This implies that the vectors $\{(\mathbf{x}_i, c) : (\mathbf{x}_i, y_i) \in \mathcal{D}^{\text{LOOM}}\}$ are linearly independent; since these are contained in $\mathbb{R}^n \times \{c\}$, which has a span of dimension $n + c$, we obtain that $k \leq n + c$. Invoking Lemma 18, we have

$$\frac{1}{\gamma} = \sum_{i=1}^m \alpha_i \geq \sum_{j=1}^k \alpha_{i_j} \geq \sum_{j=1}^k \frac{1}{(2+6c)r^2} \gamma = k \frac{1}{(2+6c)r^2} \gamma,$$

382 which implies $k \leq \frac{(2+6c)r^2}{\gamma^2}$. \square

383 6.2. Proof of the error bound

384 We are now ready for the proof of Theorem 4.

Proof of Theorem 4. Define the function $\psi : (\mathbb{R}^n)^{m+1} \rightarrow \{0, 1\}$ by

$$\begin{aligned} \psi(\mathbf{x}_1^{m+1}) &= \mathbb{1} \left[\hat{h}_m(\mathbf{x}_{m+1}; \{(\mathbf{x}_i, f^*(\mathbf{x}_i)) : i \in [m]\}) \neq f^*(\mathbf{x}_{m+1}) \right] \\ &\leq \mathbb{1} \left[(\mathbf{x}_{m+1}, f^*(\mathbf{x}_{m+1})) \in \tilde{\mathcal{D}}^{\text{LOOM}} \right], \end{aligned}$$

where $\tilde{\mathcal{D}} = \{(\mathbf{x}_i, f^*(\mathbf{x}_i)) : i \in [m+1]\}$ is determined by the arguments into ψ (the formal dependence of ψ on the values $y_i = f^*(\mathbf{x}_i)$ is suppressed, since these are determined by the \mathbf{x}_i points and the fixed target f^*). For each $t \in [m+1]$, define the permutation $\sigma_t : [m+1] \rightarrow [m+1]$ to be the one that swaps t and $m+1$ while leaving the remaining elements fixed (and in particular, σ_{m+1} is the identity map and $\sigma(\mathbf{X}_1^{m+1}) \equiv (\mathbf{X}_{\sigma_t(1)}, \dots, \mathbf{X}_{\sigma_t(m+1)})$). Since $\mathbf{X}_1, \dots, \mathbf{X}_{m+1}$ are exchangeable, and $\text{marg}(\mathcal{D}_{m+1})$ and $\max_{(\mathbf{x}, y) \in \mathcal{D}_{m+1}} \|\mathbf{x}\|$ are invariant under permutations,

$$\begin{aligned} &\mathbb{P} \left(\hat{h}_m(\mathbf{X}_{m+1}; \mathcal{D}_m) \neq Y_{m+1} \mid \gamma_{m+1}, r_{m+1} \right) \\ &= \mathbb{E} \left[\psi(\mathbf{X}_1^{m+1}) \mid \gamma_{m+1}, r_{m+1} \right] = \frac{1}{m+1} \mathbb{E} \left[\sum_{t=1}^{m+1} \psi(\sigma(\mathbf{X}_1^{m+1})) \mid \gamma_{m+1}, r_{m+1} \right]. \quad (11) \end{aligned}$$

Since for any dataset, the mistake vectors $\mathcal{D}^{\text{LOOM}}$ are invariant under permutations of \mathcal{D} , the last expression in (11) is at most $1/(m+1)$ times

$$\begin{aligned} \mathbb{E} \left[\sum_t \mathbb{1}[\mathbf{X}_{\sigma_t(m+1)} \in \mathcal{D}_{m+1}^{\text{LOOM}}] \mid \gamma_{m+1}, r_{m+1} \right] &= \mathbb{E} \left[\sum_t \mathbb{1}[\mathbf{X}_t \in \mathcal{D}_{m+1}^{\text{LOOM}}] \mid \gamma_{m+1}, r_{m+1} \right] \\ &= \mathbb{E}[\text{card}(\mathcal{D}_{m+1}^{\text{LOOM}}) \mid \gamma_{m+1}, r_{m+1}]. \end{aligned}$$

385 To show (7), we invoke Theorem 9 (recalling $\gamma_{m+1} > 0$ almost surely), to
 386 obtain $\mathbb{E}[\text{card}(\mathcal{D}_{m+1}^{\text{LOOM}}) \mid \gamma_{m+1}, r_{m+1}] \leq \min \left\{ n + c, \frac{(2+6c)r_{m+1}^2}{\gamma_{m+1}^2} \right\}$. The validity
 387 of (8) then follows by the law of total expectation and monotonicity of the
 388 expectation. \square

389 7. Transductive SVM Expected Error Bound

390 7.1. Bounding the number of pivotal vectors

391 Similarly to the above, our strategy for bounding the expected error rate of
 392 the transductive SVM is to bound the number of leave-one-out cross validation
 393 errors. In this case, however, the specification of which points correspond to
 394 such mistakes is slightly different. We refer to such points as *pivotal vectors*,
 395 and define them formally as follows.⁶

⁶As in the definition of leave-one-out mistake vectors, we also include the borderline points in this set, which suffices for our purposes of obtaining an upper bound on the number of leave-one-out prediction mistakes.

Definition 12 (Pivotal Vectors). *Given a dataset \mathcal{D} with $\gamma = \text{marg}(\mathcal{D})$, we say that $p \in [m]$ is a pivotal index if*

$$\max_{\mathbf{w} \in \mathbb{S}^n, b \in \mathbb{R}} \min_{i \in [m]} (-1)^{\mathbb{1}[i=p]} y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + cb) \geq \gamma.$$

396 *In other words, upon flipping the label of \mathbf{x}_p , the data remains linearly separable*
 397 *with margin at least γ . Let $\mathcal{D}^{\text{pivot}} \subseteq \mathcal{D}$ denote the set of all $(\mathbf{x}_p, y_p) \in \mathcal{D}$ such*
 398 *that $p \in [m]$ is a pivotal index; these are the pivotal vectors.*

399 The rest of this section is devoted to proving the following theorem, which
 400 bounds the number of pivotal vectors in terms of the dimension n and the margin
 401 γ . The proof follows the same outline as that of Theorem 9 above.

402 **Theorem 13.** *Fix any $m \in \mathbb{N}$ with $m \geq 2$, and $\mathcal{D} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_m, y_m)\}$,*
 403 *and let $(\hat{\mathbf{w}}, \hat{b}, \gamma) = \text{MMH}(\mathcal{D})$. Let $r \in \mathbb{R}_+$ be such that $\max_{i \in [m]} \|\mathbf{x}_i\| \leq r$.*
 404 *Assuming $\gamma > 0$, we have $\text{card}(\mathcal{D}^{\text{pivot}}) \leq \min \left\{ n + c, \frac{(1+3c)r^2}{\gamma^2} \right\}$.*

405 To prove this, we first note the following corollary, lower-bounding the α_p
 406 values for pivotal indices p ; it follows straightforwardly from Lemma 10.

407 **Corollary 14.** *Let \mathcal{D} , $\{\mathbf{z}_i\}$, $(\hat{\mathbf{w}}, \hat{b})$, γ , $\boldsymbol{\alpha}$, and r be as in Lemma 10. If $\gamma < \infty$,*
 408 *then $(\mathbf{x}_p, y_p) \in \mathcal{D}^{\text{pivot}} \implies \alpha_p \geq \frac{1}{(1+3c)r^2} \gamma$, $p \in [m]$.*

409 *Proof.* Suppose $\gamma < \infty$. The claim is vacuously true if $\mathcal{D}^{\text{pivot}} = \emptyset$ or $\gamma = 0$
 410 (since $\boldsymbol{\alpha} \in \mathbb{R}_+^m$), so suppose that $\mathcal{D}^{\text{pivot}}$ contains some (\mathbf{x}_p, y_p) , and that $\gamma > 0$.
 411 Also, recall from the proof of Corollary 11 that $\gamma < \infty \implies \gamma \leq r$.

412 From Definition 12, $\exists (\mathbf{w}_p, b_p) \in \mathbb{S}^n \times \mathbb{R}$ such that

$$y_p (\langle \mathbf{w}_p, \mathbf{x}_p \rangle + cb_p) \leq -\gamma$$

413 and

$$\min_{j \in [m] \setminus \{p\}} y_j (\langle \mathbf{w}_p, \mathbf{x}_j \rangle + cb_p) \geq \gamma.$$

414 Thus, since $0 < \gamma \leq r$ implies $-\gamma \in [\gamma - 4r^2/\gamma, \gamma)$, the result follows from
 415 Lemma 10 (taking $\epsilon = -\gamma$). \square

416 We are now ready for the proof of Theorem 13.

417 *Proof of Theorem 13.* If $\gamma = \infty$ (which can only happen if $c = 1$), then it
 418 must be that every $(\mathbf{x}_i, y_i) \in \mathcal{D}$ has the same y_i . Since $m \geq 2$, flipping any
 419 label yields a data set with at least one of each label, and hence finite margin:
 420 that is, $\forall i \in [m]$, $\text{marg}((\mathcal{D} \setminus \{(\mathbf{x}_i, y_i)\}) \cup \{(\mathbf{x}_i, -y_i)\}) < \infty = \gamma$. Thus, if
 421 $\gamma = \infty$, $\mathcal{D}^{\text{pivot}} = \emptyset$, so that the result trivially follows in this case. Furthermore,
 422 note that if $r = 0$ (which again can only happen if $c = 1$, due to the $\gamma > 0$
 423 assumption), then every $(\mathbf{x}_i, y_i) \in \mathcal{D}$ has the same \mathbf{x}_i . Together with the linear
 424 separability assumption, this again implies that every $(\mathbf{x}_i, y_i) \in \mathcal{D}$ has the same

425 y_i , so that $\gamma = \infty$, and hence, as just established, the result trivially holds in
 426 this case.

For the remaining case, suppose $0 < \gamma < \infty$ and $r > 0$, and put $k = \text{card}(\mathcal{D}^{\text{pivot}})$. The claim is trivial if $k = 0$, so assume $k \geq 1$ and let $\{i_1, \dots, i_k\} \subseteq [m]$ be the pivotal indices: $\mathcal{D}^{\text{pivot}} = \{(\mathbf{x}_{i_1}, y_{i_1}), \dots, (\mathbf{x}_{i_k}, y_{i_k})\}$. Put $\mathbf{z}_i = y_i \mathbf{x}_i$, $i \in [m]$. Lemma 17 implies the existence of $\boldsymbol{\alpha} \in \mathbb{R}_+^m$ satisfying the conditions (i,ii,iii) of Lemma 7, such that the vectors $\{(\mathbf{x}_i, c) : i \in \text{supp}(\boldsymbol{\alpha})\}$ are linearly independent. Furthermore, Corollary 14 implies that for any such $\boldsymbol{\alpha}$, $\alpha_{i_j} \geq \frac{1}{(1+3c)r^2} \gamma > 0$, $j \in [k]$. Thus, any pivotal index i_j must be in $\text{supp}(\boldsymbol{\alpha})$, and hence $\{(\mathbf{x}_i, c) : (\mathbf{x}_i, y_i) \in \mathcal{D}^{\text{pivot}}\} \subseteq \{(\mathbf{x}_i, c) : i \in \text{supp}(\boldsymbol{\alpha})\}$. This implies that the vectors $\{(\mathbf{x}_i, c) : (\mathbf{x}_i, y_i) \in \mathcal{D}^{\text{pivot}}\}$ are linearly independent; since these are contained in $\mathbb{R}^n \times \{c\}$, which has a span of dimension $n + c$, we obtain that $k \leq n + c$. Invoking Lemma 18, we have

$$\frac{1}{\gamma} = \sum_{i=1}^m \alpha_i \geq \sum_{j=1}^k \alpha_{i_j} \geq \sum_{j=1}^k \frac{1}{(1+3c)r^2} \gamma = k \frac{1}{(1+3c)r^2} \gamma,$$

427 which implies $k \leq \frac{(1+3c)r^2}{\gamma^2}$. □

428 7.2. Proof of the transductive SVM error bound

429 We are now ready for the proof of Theorem 5.

Proof of Theorem 5. The proof is nearly identical to that of Theorem 4, except based on pivotal vectors instead of leave-one-out mistake vectors. Define the function $\psi : (\mathbb{R}^n)^{m+1} \rightarrow \{0, 1\}$ by

$$\begin{aligned} \psi(\mathbf{x}_1, \dots, \mathbf{x}_{m+1}) &= \mathbb{1} \left[\hat{h}_m(\mathbf{x}_{m+1}; \{(\mathbf{x}_i, f^*(\mathbf{x}_i)) : i \in [m]\}) \neq f^*(\mathbf{x}_{m+1}) \right] \\ &\leq \mathbb{1} \left[(\mathbf{x}_{m+1}, f^*(\mathbf{x}_{m+1})) \in \tilde{\mathcal{D}}^{\text{pivot}} \right], \end{aligned}$$

430 where $\tilde{\mathcal{D}} = \{(\mathbf{x}_i, f^*(\mathbf{x}_i)) : i \in [m+1]\}$ is determined by the arguments into ψ .
 431 Define the permutation $\sigma_t : [m+1] \rightarrow [m+1]$, $t \in [m+1]$, as in the proof of
 432 Theorem 4.

Since $\mathbf{X}_1, \dots, \mathbf{X}_{m+1}$ are exchangeable random variables, and both $\text{marg}(\mathcal{D}_{m+1})$ and $\max_{(\mathbf{x}, y) \in \mathcal{D}_{m+1}} \|\mathbf{x}\|$ are invariant under permutations,

$$\begin{aligned} \mathbb{P} \left(\hat{h}_m(\mathbf{X}_{m+1}; \mathcal{D}_m) \neq Y_{m+1} \mid \gamma_{m+1}, r_{m+1} \right) &= \mathbb{E} [\psi(\mathbf{X}_1, \dots, \mathbf{X}_{m+1}) \mid \gamma_{m+1}, r_{m+1}] \\ &= \frac{1}{m+1} \sum_{t=1}^{m+1} \mathbb{E} [\psi(\mathbf{X}_{\sigma_t(1)}, \dots, \mathbf{X}_{\sigma_t(m+1)}) \mid \gamma_{m+1}, r_{m+1}] \\ &= \frac{1}{m+1} \mathbb{E} \left[\sum_{t=1}^{m+1} \psi(\mathbf{X}_{\sigma_t(1)}, \dots, \mathbf{X}_{\sigma_t(m+1)}) \mid \gamma_{m+1}, r_{m+1} \right]. \end{aligned} \quad (12)$$

Since for any dataset, the pivotal vectors $\mathcal{D}^{\text{pivot}}$ are invariant under permutations of \mathcal{D} , the last expression in (12) is at most

$$\begin{aligned} & \frac{1}{m+1} \mathbb{E} \left[\sum_{t=1}^{m+1} \mathbb{1} \left[\mathbf{X}_{\sigma_t(m+1)} \in \mathcal{D}_{m+1}^{\text{pivot}} \right] \middle| \gamma_{m+1}, r_{m+1} \right] \\ &= \frac{1}{m+1} \mathbb{E} \left[\sum_{t=1}^{m+1} \mathbb{1} \left[\mathbf{X}_t \in \mathcal{D}_{m+1}^{\text{pivot}} \right] \middle| \gamma_{m+1}, r_{m+1} \right] \\ &= \frac{1}{m+1} \mathbb{E} \left[\text{card} \left(\mathcal{D}_{m+1}^{\text{pivot}} \right) \middle| \gamma_{m+1}, r_{m+1} \right]. \end{aligned}$$

To show (9), we invoke Theorem 13 (recalling that $\gamma_{m+1} > 0$ almost surely) to obtain

$$\mathbb{E} \left[\text{card} \left(\mathcal{D}_{m+1}^{\text{pivot}} \right) \middle| \gamma_{m+1}, r_{m+1} \right] \leq \min \left\{ n + c, \frac{(1 + 3c)r_{m+1}^2}{\gamma_{m+1}^2} \right\}.$$

433 The validity of (10) then follows by the law of total expectation and monotonicity of the expectation. \square

435 *Remark.* The above argument can equivalently be interpreted as arguing that
436 transductive SVM corresponds to predicting with the one-inclusion graph prediction
437 strategy of [2], with an orientation of the graph having out-degree of the
438 target node at most $\min\{n + c, (1 + 3c)/\gamma_{m+1}^2\}$.

439 8. Agnostic Case

440 Here we extend the results above to the agnostic case. In this case, there is a
441 distribution P_{XY} over $\mathbb{R}^n \times \{-1, 1\}$, the data (\mathbf{X}_i, Y_i) are i.i.d. P_{XY} -distributed
442 samples, and the error rate $\text{err}(h)$ of a classifier h is defined as $\mathbb{P}(h(\mathbf{X}) \neq Y)$
443 for $(\mathbf{X}, Y) \sim P_{XY}$. Again, the advantage of the results here over the standard
444 treatment in textbooks is the explicit handling of the nonhomogeneous case. As
445 discussed above, this explicit treatment of the bias term can dramatically improve
446 the bounds compared to the naïve approach of adding an extra dimension and
447 bounding the risk in terms of the homogeneous-case margin bounds in the
448 resulting $n + 1$ dimensional problem: specifically, improving the dependence on
449 R , the magnitude of the data.

In the agnostic case, the support vector machine corresponds to the following optimization problem.

$$\begin{aligned} & \text{minimize} && \|w\|^2 + C \sum_{i=1}^m \xi_i \\ & \text{subject to} && Y_i(\langle \mathbf{w}, \mathbf{X}_i \rangle + b) \geq 1 - \xi_i, \forall i \leq m \\ & && \xi_i \geq 0, \forall i \leq m. \end{aligned}$$

450 We are therefore interested in expressing the generalization bound in terms of
451 $\|w\|^2$ and $\sum_{i=1}^m \xi_i$ at the solution. In particular, we have the following theorem.

Theorem 15. Let $(\hat{w}, \hat{b}, \hat{\xi})$ denote the values at the solution of the above optimization problem, and let \hat{h}_m denote the resulting classifier $\mathbf{x} \mapsto \text{sign}(\langle \hat{\mathbf{w}}, \mathbf{x} \rangle + \hat{b})$. Then with probability at least $1 - \delta$, letting $R = \max_{i \in [m]} \|\mathbf{X}_i\|$,

$$\text{err}(\hat{h}_m) \leq \frac{1}{m} \sum_{i=1}^m \hat{\xi}_i + 4\sqrt{\frac{([\![R]\!] \|\hat{w}\| + 1)^2}{m}} + 3\sqrt{\frac{\ln\left(\frac{\pi^4 [\![R]\!]^2 [\|\hat{w}\|]^2}{18\delta}\right)}{m}}.$$

452 *Proof.* The proof of this follows a standard argument (see e.g., [4]), with a few
 453 modifications to explicitly account for the bias term (which does not appear in
 454 the bound). First, for any $\mathbf{w} \in \mathbb{R}^n$ and $b \in \mathbb{R}$, for $(\mathbf{x}, y) \in \mathbb{R}^n \times \{-1, 1\}$, define
 455 $h_{\mathbf{w}, b}(\mathbf{x}) = \text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle + b)$ and $\ell_{\mathbf{w}, b}(\mathbf{x}, y) = \min\{\max\{1 - y(\langle \mathbf{w}, \mathbf{x} \rangle + b), 0\}, 1\}$.
 456 Then define $H_\Lambda = \{\ell_{\mathbf{w}, b} : \|\mathbf{w}\| \leq \Lambda, b \in \mathbb{R}\}$ for any $\Lambda > 0$. Now we note that,
 457 for any \mathbf{w}, b , $\text{err}(h_{\mathbf{w}, b}) \leq \mathbb{E}[\ell_{\mathbf{w}, b}(\mathbf{X}, Y)]$ for $(\mathbf{X}, Y) \sim P_{XY}$. Thus, it suffices to
 458 bound $\mathbb{E}[\ell_{\hat{\mathbf{w}}, \hat{b}}(\mathbf{X}, Y) | \hat{\mathbf{w}}, \hat{b}]$.

Fix any $\Lambda, R > 0$. Theorem 3.1 of [4] implies that, with probability at least $1 - \delta'$, every $\ell_{\mathbf{w}, b} \in H_\Lambda$ satisfies

$$\mathbb{E}[\ell_{\mathbf{w}, b}(\mathbf{X}, Y)] \leq \frac{1}{m} \sum_{i=1}^m \ell_{\mathbf{w}, b}(\mathbf{X}_i, Y_i) + 2\text{Rademacher}(H_\Lambda) + 3\sqrt{\frac{\ln\left(\frac{2}{\delta'}\right)}{m}},$$

where

$$\text{Rademacher}(H_\Lambda) = \mathbb{E} \left[\sup_{f_{\mathbf{w}, b} \in H_\Lambda} \frac{1}{m} \sum_{i=1}^m \epsilon_i f_{\mathbf{w}, b}(\mathbf{X}_i, Y_i) \middle| \{(\mathbf{X}_i, Y_i)\}_{i \in [m]} \right]$$

and $\epsilon_1, \dots, \epsilon_m$ are independent Uniform($\{-1, 1\}$) random variables, independent from $\{(\mathbf{X}_i, Y_i)\}_{i \in [m]}$. Now note that, if $\max_{i \in [m]} \|\mathbf{X}_i\| \leq R$, then for any $\mathbf{w} \in \mathbb{R}^n$ with $\|\mathbf{w}\| \leq \Lambda$, for any $b > R\Lambda + 1$, $\ell_{\mathbf{w}, b}(\mathbf{X}_i, Y_i) = \ell_{\mathbf{w}, R\Lambda+1}(\mathbf{X}_i, Y_i)$, and for any $b < -(R\Lambda + 1)$, $\ell_{\mathbf{w}, b}(\mathbf{X}_i, Y_i) = \ell_{\mathbf{w}, -(R\Lambda+1)}(\mathbf{X}_i, Y_i)$. Thus, when $\max_{i \in [m]} \|\mathbf{X}_i\| \leq R$, H_Λ can equivalently be defined as $\{\ell_{\mathbf{w}, b} : \|\mathbf{w}\| \leq \Lambda, |b| \leq R\Lambda + 1\}$. Also note that the function $\ell_{\mathbf{w}, b}(\mathbf{x}, y)$ is 1-Lipschitz in $(\mathbf{x}, y) \mapsto y(\langle \mathbf{w}, \mathbf{x} \rangle + b)$. Combining these two facts with Lemma 4.2 of [4] implies $\text{Rademacher}(H_\Lambda)$ is at most

$$\begin{aligned} & \mathbb{E} \left[\sup_{\mathbf{w}, b: \|\mathbf{w}\| \leq \Lambda, |b| \leq R\Lambda+1} \frac{1}{m} \sum_{i=1}^m \epsilon_i Y_i (\langle \mathbf{w}, \mathbf{X}_i \rangle + b) \middle| \{(\mathbf{X}_i, Y_i)\}_{i \in [m]} \right] \\ &= \mathbb{E} \left[\sup_{\mathbf{w}, b: \|\mathbf{w}\| \leq \Lambda, |b| \leq R\Lambda+1} \frac{1}{m} \sum_{i=1}^m \epsilon_i (\langle \mathbf{w}, \mathbf{X}_i \rangle + b) \middle| \{\mathbf{X}_i\}_{i \in [m]} \right] \\ &= \frac{1}{m} \mathbb{E} \left[\sup_{\mathbf{w}: \|\mathbf{w}\| \leq \Lambda} \left\langle \mathbf{w}, \sum_{i=1}^m \epsilon_i \mathbf{X}_i \right\rangle + \sup_{b: |b| \leq R\Lambda+1} b \sum_{i=1}^m \epsilon_i \middle| \{\mathbf{X}_i\}_{i \in [m]} \right] \\ &= \frac{1}{m} \mathbb{E} \left[\sup_{\mathbf{w}: \|\mathbf{w}\| \leq \Lambda} \left\langle \mathbf{w}, \sum_{i=1}^m \epsilon_i \mathbf{X}_i \right\rangle \middle| \{\mathbf{X}_i\}_{i \in [m]} \right] + \frac{R\Lambda + 1}{m} \mathbb{E} \left[\left| \sum_{i=1}^m \epsilon_i \right| \right] \\ &\leq \frac{1}{m} \mathbb{E} \left[\Lambda \left\| \sum_{i=1}^m \epsilon_i \mathbf{X}_i \right\| \middle| \{\mathbf{X}_i\}_{i \in [m]} \right] + \frac{R\Lambda + 1}{m} \mathbb{E} \left[\left| \sum_{i=1}^m \epsilon_i \right| \right]. \end{aligned}$$

Jensen's inequality implies this is at most

$$\frac{\Lambda}{m} \mathbb{E} \left[\left\| \sum_{i=1}^m \epsilon_i \mathbf{X}_i \right\|^2 \middle| \{\mathbf{X}_i\}_{i \in [m]} \right]^{1/2} + \frac{R\Lambda + 1}{m} \mathbb{E} \left[\left| \sum_{i=1}^m \epsilon_i \right|^2 \right]^{1/2},$$

and the fact that the ϵ_i variables have zero mean and are independent implies this is equal

$$\begin{aligned} & \frac{\Lambda}{m} \mathbb{E} \left[\sum_{i=1}^m \epsilon_i^2 \|\mathbf{X}_i\|^2 \middle| \{\mathbf{X}_i\}_{i \in [m]} \right]^{1/2} + \frac{R\Lambda + 1}{m} \mathbb{E} \left[\sum_{i=1}^m \epsilon_i^2 \right]^{1/2} \\ &= \frac{\Lambda}{m} \left(\sum_{i=1}^m \epsilon_i^2 \|\mathbf{X}_i\|^2 \right)^{1/2} + \frac{R\Lambda + 1}{m} \sqrt{m} \leq \frac{\Lambda}{m} \sqrt{mR^2} + \frac{R\Lambda + 1}{\sqrt{m}} \leq 2\sqrt{\frac{(R\Lambda + 1)^2}{m}}. \end{aligned}$$

Thus, for any $\delta_{R,\Lambda} \in (0, 1)$, with probability at least $1 - \delta_{R,\Lambda}$, if $\max_{i \in [m]} \|\mathbf{X}_i\| \leq R$, then every $(\mathbf{w}, b) \in \mathbb{R}^n \times \mathbb{R}$ with $\|\mathbf{w}\| \leq \Lambda$ satisfies

$$\text{err}(h_{\mathbf{w},b}) \leq \frac{1}{m} \sum_{i=1}^m \ell_{\mathbf{w},b}(\mathbf{X}_i, Y_i) + 4\sqrt{\frac{(R\Lambda + 1)^2}{m}} + 3\sqrt{\frac{\ln\left(\frac{2}{\delta_{R,\Lambda}}\right)}{m}}.$$

Now let $p_i = q_i = 6/(\pi i)^2$ for $i \in \mathbb{N}$, and define $\delta_{R,\Lambda} = p_R q_\Lambda \delta$ for $R, \Lambda \in \mathbb{N}$. Then by a union bound, with probability at least $1 - \sum_{R \in \mathbb{N}} \sum_{\Lambda \in \mathbb{N}} \delta_{R,\Lambda} = 1 - \delta$, the above claim holds simultaneously for all $R, \Lambda \in \mathbb{N}$. In particular, on this event, taking $R = \lceil \max_{i \in [m]} \|\mathbf{X}_i\| \rceil$ and $\Lambda = \lceil \|\hat{\mathbf{w}}\| \rceil$, we have

$$\text{err}(h_{\hat{\mathbf{w}},\hat{b}}) \leq \frac{1}{m} \sum_{i=1}^m \ell_{\hat{\mathbf{w}},\hat{b}}(\mathbf{X}_i, Y_i) + 4\sqrt{\frac{(R\Lambda + 1)^2}{m}} + 3\sqrt{\frac{\ln\left(\frac{\pi^4 R^2 \Lambda^2}{18\delta}\right)}{m}}.$$

459 The result then follows from this by noting that $\hat{h}_m = h_{\hat{\mathbf{w}},\hat{b}}$, and that $\ell_{\hat{\mathbf{w}},\hat{b}}(\mathbf{X}_i, Y_i)$
 460 $\leq \max\{1 - Y_i(\langle \hat{\mathbf{w}}, \mathbf{X}_i \rangle + \hat{b}), 0\}$, and by the constraints in the optimization problem,
 461 we know $\hat{\xi}_i \geq \max\{1 - Y_i(\langle \hat{\mathbf{w}}, \mathbf{X}_i \rangle + \hat{b}), 0\}$, so that $\frac{1}{m} \sum_{i=1}^m \ell_{\hat{\mathbf{w}},\hat{b}}(\mathbf{X}_i, Y_i) \leq$
 462 $\frac{1}{m} \sum_{i=1}^m \hat{\xi}_i$. \square

463 [1] V. Vapnik, O. Chapelle, Bounds on error expectation for support vector
 464 machines, *Neural Computation* 12 (9) (2000) 2013–2036.

465 [2] D. Haussler, N. Littlestone, M. K. Warmuth,
 466 Predicting $\{0,1\}$ -functions on randomly drawn points, *Inf. Comput.*
 467 115 (2) (1994) 248–292. doi:10.1006/inco.1994.1097.
 468 URL <http://dx.doi.org/10.1006/inco.1994.1097>

469 [3] L. Devroye, L. Györfi, G. Lugosi, A probabilistic theory of pattern recognition,
 470 Vol. 31 of *Applications of Mathematics* (New York), Springer-Verlag,
 471 New York, 1996.

- 472 [4] M. Mohri, A. Rostamizadeh, A. Talwalkar, Foundations Of Machine Learning,
473 The MIT Press, 2012.
- 474 [5] A. Ehrenfeucht, D. Haussler, M. Kearns, L. Valiant, A general lower bound
475 on the number of examples needed for learning, Information and Compu-
476 tation 82 (1989) 247–261.
- 477 [6] R. Herbrich, Learning Kernel Classifiers: Theory and Algorithms, 2nd Edi-
478 tion, The MIT Press, 2002.
- 479 [7] J. Shawe-Taylor, Private communication (2015).
- 480 [8] N. Littlestone, From on-line to batch learning, in: Proceedings of the Sec-
481 ond Annual Workshop on Computational Learning Theory, COLT '89,
482 Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1989, pp.
483 269–284.
484 URL <http://dl.acm.org/citation.cfm?id=93335.93365>
- 485 [9] A. Novikoff, On convergence proofs for perceptrons, in: Proc. Sympos.
486 Math. Theory of Automata (New York, 1962), Polytechnic Press of Poly-
487 technic Inst. of Brooklyn, Brooklyn, N.Y., 1963, pp. 615–622.
- 488 [10] T. Zhang, Covering number bounds of certain regularized linear function
489 classes, The Journal of Machine Learning Research 2 (2002) 527–550.
- 490 [11] C. J. C. Burges, A tutorial on support vector machines for pattern recog-
491 nition, Data Min. Knowl. Discov. 2 (2) (1998) 121–167.
- 492 [12] N. Cristianini, J. Shawe-Taylor, An introduction to support vector ma-
493 chines and other kernel-based learning methods, Cambridge University
494 Press, 2000.
- 495 [13] J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, M. Anthony, Structural
496 risk minimization over data-dependent hierarchies, IEEE Transactions on
497 Information Theory 44 (5) (1998) 1926–1940.
- 498 [14] J. M. Borwein, A. S. Lewis, Convex analysis and nonlinear optimization,
499 2nd Edition, CMS Books in Mathematics/Ouvrages de Mathématiques
500 de la SMC, 3, Springer, New York, 2006, theory and examples.
501 doi:10.1007/978-0-387-31256-9.
502 URL <http://dx.doi.org/10.1007/978-0-387-31256-9>
- 503 [15] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University
504 Press, 2004.
- 505 [16] B. Schölkopf, A. J. Smola, Learning with Kernels: Support Vector Ma-
506 chines, Regularization, Optimization, and Beyond, The MIT Press, 2002.
- 507 [17] V. N. Vapnik, Statistical Learning Theory, Wiley-Interscience, 1998.

508 **Appendix A. Technical lemmas**

Lemma 16. $\forall \mathbf{u} \in \mathbb{R}^n, \forall \mathbf{v}, \mathbf{w} \in \mathbb{S}^n, \forall \delta > 0,$

$$|\langle \mathbf{u}, \mathbf{v} \rangle - \langle \mathbf{u}, \mathbf{w} \rangle| \geq \delta \implies \langle \mathbf{v}, \mathbf{w} \rangle \leq 1 - \frac{\delta^2}{2\|\mathbf{u}\|^2}.$$

Proof. Suppose $|\langle \mathbf{u}, \mathbf{v} \rangle - \langle \mathbf{u}, \mathbf{w} \rangle| \geq \delta > 0$. First note that, in particular, this implies $\|\mathbf{u}\| > 0$, so that $\frac{\delta^2}{2\|\mathbf{u}\|^2}$ is well-defined. Denote $\sigma = \|\mathbf{u}\|$. By rotational symmetry, there is no loss of generality in assuming $\mathbf{u} = \sigma \mathbf{e}_1$, where \mathbf{e}_1 denotes the first canonical orthonormal basis vector. Hence, $\langle \mathbf{u}, \mathbf{v} \rangle = \sigma v_1$ and $\langle \mathbf{u}, \mathbf{w} \rangle = \sigma w_1$, so that we have that

$$|v_1 - w_1| \geq \delta/\sigma. \tag{A.1}$$

Furthermore, since $\|\mathbf{v}\| = \|\mathbf{w}\| = 1$, the Cauchy-Schwarz inequality implies

$$\langle \mathbf{v}, \mathbf{w} \rangle \leq v_1 w_1 + \sqrt{(1 - v_1^2)(1 - w_1^2)}.$$

509 It remains to show that the right-hand side of the above display is at most
 510 $1 - (v_1 - w_1)^2/2$; together with (A.1), this will imply $\langle \mathbf{v}, \mathbf{w} \rangle \leq 1 - \frac{\delta^2}{2\sigma^2}$, as
 511 claimed. To this end, we claim that

$$st + \sqrt{(1 - s^2)(1 - t^2)} \leq 1 - \frac{1}{2}(s - t)^2, \quad 0 \leq s, t \leq 1. \tag{A.2}$$

Put $L = \sqrt{(1 - s^2)(1 - t^2)}$ and $R = 1 - (s - t)^2/2 - st$; clearly, (A.2) is equivalent to the assertion that $L^2 \leq R^2$. Now

$$R^2 - L^2 = \frac{(s^2 - t^2)^2}{4} \geq 0.$$

512 This proves (A.2). □

513 **Appendix B. Facts about support vectors**

514 The next two results are known, but we reprove them here to establish them
 515 in the particular form we require for their use in the proofs of our other results.

516 **Lemma 17.** *Let $\mathcal{D}, \{\mathbf{z}_i\}, (\hat{\mathbf{w}}, \hat{b})$, and γ be as in Lemma 7. Then the $\boldsymbol{\alpha} \in \mathbb{R}_+^m$
 517 in Lemma 7(i,ii,iii) may be chosen so that the vectors $\{(\mathbf{x}_i, c) : i \in \text{supp}(\boldsymbol{\alpha})\}$
 518 are linearly independent.*

519 **Remark:** Obviously, whenever $(\hat{\mathbf{w}}, 0) \in \text{span}(\{(\mathbf{z}_i, cy_i)\})$, there exist linearly
 520 independent $\{(\mathbf{z}'_i, cy'_i)\} \subseteq \{(\mathbf{z}_i, cy_i)\}$ such that $(\hat{\mathbf{w}}, 0) \in \text{span}(\{(\mathbf{z}'_i, cy'_i)\})$. What
 521 makes the claim nontrivial is the extra condition of nonnegativity on $\boldsymbol{\alpha}$.

Proof. This argument is essentially taken from [17]. Let $\boldsymbol{\alpha} \in \mathbb{R}_+^m$ be such that $\|\boldsymbol{\alpha}\|_0$ is minimal, subject to the conditions in Lemma 7(i,ii,iii). Put $k = \|\boldsymbol{\alpha}\|_0$ and let $\{i_1, \dots, i_k\} = \text{supp}(\boldsymbol{\alpha})$. For the sake of obtaining a contradiction, suppose the vectors $(\mathbf{x}_{i_1}, c), \dots, (\mathbf{x}_{i_k}, c)$ are not linearly independent. This implies that $(\mathbf{z}_{i_1}, y_{i_1}c), \dots, (\mathbf{z}_{i_k}, y_{i_k}c)$ are also not linearly independent. Thus, there exist scalars $\beta_{i_\ell} \in \mathbb{R}$, $\ell \in [k]$, not all equal zero, such that

$$\sum_{\ell \in [k]} \beta_{i_\ell} (\mathbf{z}_{i_\ell}, y_{i_\ell}c) = \mathbf{0},$$

522 where here $\mathbf{0}$ is the $(n+1)$ -dimensional vector of 0's. Now for each $t \in \mathbb{R}$, define
 523 $\boldsymbol{\alpha}^{(t)} \in \mathbb{R}^m$ as

$$\alpha_i^{(t)} = \begin{cases} 0, & i \notin \{i_1, \dots, i_k\} \\ \alpha_i - t\beta_{i_\ell}, & i \in \{i_1, \dots, i_k\} \end{cases}.$$

524 Then

$$\sum_{i=1}^m \alpha_i^{(t)} (\mathbf{z}_i, y_i c) = \sum_{\ell \in [k]} \alpha_{i_\ell} (\mathbf{z}_{i_\ell}, y_{i_\ell} c) - t \sum_{\ell \in [k]} \beta_{i_\ell} (\mathbf{z}_{i_\ell}, y_{i_\ell} c) = (\hat{\mathbf{w}}, 0) - \mathbf{0} = (\hat{\mathbf{w}}, 0),$$

525 so that $\boldsymbol{\alpha}^{(t)}$ also satisfies the conditions (i,iii) of Lemma 7, aside from the non-
 526 negativity requirement ($\boldsymbol{\alpha}^{(t)} \in \mathbb{R}_+^m$). Furthermore, any $i \in [m]$ with $\alpha_i^{(t)} \neq 0$
 527 has $i \in \{i_1, \dots, i_k\}$, so that $\alpha_i > 0$, and hence condition (ii) of Lemma 7 implies
 528 $(\mathbf{x}_i, y_i) \in \mathcal{D}^{\text{marg}}$; therefore, $\boldsymbol{\alpha}^{(t)}$ also satisfies condition (ii) of Lemma 7.

529 Next, for each $\ell \in [k]$ with $\beta_{i_\ell} \neq 0$ (of which there is at least one), define
 530 $t_\ell = \alpha_{i_\ell} / \beta_{i_\ell}$. Since each α_{i_ℓ} is strictly greater than 0, and β_{i_ℓ} is finite, each
 531 of these values t_ℓ is a nonzero finite value. Let t^* denote the value t_{ℓ^*} for the
 532 value $\ell^* \in [k]$ with smallest $|t_\ell|$ among $\ell \in [k]$ with $\beta_{i_\ell} \neq 0$. Then note that
 533 every $\ell \in [k]$ has $\alpha_{i_\ell}^{(t^*)} = \alpha_{i_\ell} - t^* \beta_{i_\ell} \geq 0$, so that $\boldsymbol{\alpha}^{(t^*)} \in \mathbb{R}_+^m$. Furthermore,
 534 $\alpha_{i_{\ell^*}}^{(t^*)} = \alpha_{i_{\ell^*}} - t_{\ell^*} \beta_{i_{\ell^*}} = 0$. Thus, $\boldsymbol{\alpha}^{(t^*)} \in \mathbb{R}_+^m$. However, $\|\boldsymbol{\alpha}^{(t^*)}\|_0 \leq \|\boldsymbol{\alpha}\|_0 - 1$.
 535 Altogether, we have that $\boldsymbol{\alpha}^{(t^*)} \in \mathbb{R}_+^m$ satisfies the conditions (i,ii,iii) of Lemma 7,
 536 while $\|\boldsymbol{\alpha}^{(t^*)}\|_0 < \|\boldsymbol{\alpha}\|_0$. This violates the minimality of $\|\boldsymbol{\alpha}\|_0$ stipulated in our
 537 choice of $\boldsymbol{\alpha}$, resulting in a contradiction. We therefore conclude that, for any
 538 $\boldsymbol{\alpha} \in \mathbb{R}_+^m$ with minimal $\|\boldsymbol{\alpha}\|_0$ subject to the constraints in Lemma 7(i,ii,iii), the
 539 vectors $\{(\mathbf{x}_i, c) : i \in \text{supp}(\boldsymbol{\alpha})\}$ are linearly independent. Since the existence of
 540 such $\boldsymbol{\alpha}$ is guaranteed by Lemma 7(i,ii,iii) (and the fact that $\|\boldsymbol{\alpha}\|_0$ can take only
 541 finitely many different values), the result follows. \square

542 The following result establishes a connection between the Lagrange multi-
 543 pliers $\boldsymbol{\alpha}$ and the margin γ . The result is well known, but we include a proof
 544 (taken from [4]) for completeness, and since our definitions are slightly different
 545 (in the normalization).

Lemma 18. Let $\{\mathbf{z}_i\}$, $(\hat{\mathbf{w}}, \hat{b})$, and γ be as in Lemma 7, with $\boldsymbol{\alpha} \in \mathbb{R}_+^m$ satisfying (i,ii,iii) therein. Then

$$\sum_{i=1}^m \alpha_i = \frac{1}{\gamma}.$$

Proof. This proof is taken from [4]. Conditions (i,ii) of Lemma 7 imply that, for any $i \in \text{supp}(\boldsymbol{\alpha})$,

$$\hat{c}\hat{b} + \sum_{j=1}^m \alpha_j y_j \langle \mathbf{x}_j, \mathbf{x}_i \rangle = \hat{c}\hat{b} + \langle \hat{\mathbf{w}}, \mathbf{x}_i \rangle + \hat{c}\hat{b} = y_i \gamma.$$

Multiplying by $\alpha_i y_i$, we have

$$\hat{c}\hat{b}\alpha_i y_i + \sum_{j=1}^m \alpha_j \alpha_i y_j y_i \langle \mathbf{x}_j, \mathbf{x}_i \rangle = \alpha_i y_i^2 \gamma = \alpha_i \gamma.$$

Furthermore, this is trivially also satisfied for any $i \notin \text{supp}(\boldsymbol{\alpha})$, since the expressions are all equal zero in that case. Thus, summing over all $i \in [m]$, we obtain

$$\hat{b}\hat{c} \sum_{i=1}^m \alpha_i y_i + \sum_{i=1}^m \sum_{j=1}^m \alpha_j \alpha_i y_j y_i \langle \mathbf{x}_j, \mathbf{x}_i \rangle = \gamma \sum_{i=1}^m \alpha_i.$$

546 Conditions (i,iii) of Lemma 7 imply that the left hand side of the above equals
 547 $0 + \langle \hat{\mathbf{w}}, \hat{\mathbf{w}} \rangle = 1$, so that $\gamma \sum_{i=1}^m \alpha_i = 1$, or equivalently, $\sum_{i=1}^m \alpha_i = \frac{1}{\gamma}$. \square

548 Appendix C. Lower Bounds

549 Here we sketch a proof of the lower bound (1). In particular, combined with
 550 the above upper bounds, this establishes that the support vector machine (in
 551 both the inductive and transductive variant) achieves the minimax expected
 552 error rate in the limit, up to constant factors.

Theorem 19. For any learning algorithm A , there exists a data distribution and target function such that the maximum margin homogeneous linear separator for m samples has margin at least γ (almost surely), and the expected error rate of A (with these m samples as input) is at least

$$\frac{\min\{1/\gamma^2, n\} - 1}{2e(m+1)}.$$

553 *Proof Sketch.* It was proven in [2] that, for any space \mathcal{X} and any concept space
 554 \mathcal{H} of a given VC dimension d , there exists a distribution on \mathcal{X} such that, for
 555 any learning algorithm A , there exists a choice of target function in \mathcal{H} such that
 556 the expected error rate of A is at least $(d-1)/(2e(m+1))$, given m iid samples.
 557 Furthermore, the distribution of the data in that proof can be supported on
 558 an arbitrary shatterable set of size d . We establish our result by reduction to

559 this one. Specifically, we note that the first $k = \min\{1/\gamma^2, n\}$ basis vectors
560 are shatterable by homogeneous linear separators having margin at least γ with
561 respect to these k points. Thus, restricting to a concept space of 2^k homogeneous
562 linear separators with margin at least γ on these k points, the VC dimension
563 is k , which establishes a lower bound $(k - 1)/(2e(m + 1))$ for this subspace.
564 Since these separators are contained in the larger space of all linear separators,
565 and the lower bound also applies to improper learning algorithms, this lower
566 bound also holds for the full space of linear separators. Furthermore, we have
567 established this lower bound while restricting the target concept to be among
568 these 2^k separators, each of which has margin at least γ on the points in the
569 support of the data distribution, and therefore (almost surely) has margin at
570 least γ on the m data points. \square