On the proliferation of support vectors in high dimensions

Daniel Hsu
Columbia University

Vidya Muthukumar
Georgia Institute of Technology

Ji Xu
Columbia University

Abstract

The support vector machine (SVM) is a well-established classification method whose name refers to the particular training examples, called support vectors, that determine the maximum margin separating hyperplane. The SVM classifier is known to enjoy good generalization properties when the number of support vectors is small compared to the number of training examples. However, recent research has shown that in sufficiently high-dimensional linear classification problems, the SVM can generalize well despite a *proliferation of support vectors* where all training examples are support vectors. In this paper, we identify new deterministic equivalences for this phenomenon of support vector proliferation, and use them to (1) substantially broaden the conditions under which the phenomenon occurs in high-dimensional settings, and (2) prove a nearly matching converse result.

1 INTRODUCTION

The Support Vector Machine (SVM) is one of the most well-known and commonly used methods for binary classification in machine learning (Vapnik, 1982; Cortes and Vapnik, 1995). Its homogeneous version in the linearly separable setting (commonly also known as the *hard-margin SVM*) is defined as the solution to an optimization problem characterizing the linear classifier (a separating hyperplane) that maximizes the minimum margin achieved on the n training examples $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n) \in \mathbb{R}^d \times \{-1, +1\}$:

$$\begin{aligned} \max_{\boldsymbol{w} \in \mathbb{R}^d, \gamma \geq 0} \quad & \gamma \\ \text{subj. to} \quad & \mathrm{margin}_i(\boldsymbol{w}) \geq \gamma \quad \text{for all } i = 1, \ldots, n, \end{aligned} \tag{1}$$

where

$$\mathrm{margin}_i(\boldsymbol{w}) := \begin{cases} y_i \boldsymbol{x}_i^\intercal \boldsymbol{w} / \|\boldsymbol{w}\|_2 & \text{if } \boldsymbol{w} \neq \boldsymbol{0} \\ 0 & \text{if } \boldsymbol{w} = \boldsymbol{0} \end{cases}$$

is the margin achieved by $\boldsymbol{w}$ on the i^{th} training example[1] $(\boldsymbol{x}_i, y_i)$. The SVM gets its name from the fact that the solution $(\boldsymbol{w}^\star, \gamma^\star)$ depends only on the set of training examples that achieve the minimum margin value, $\gamma^\star$. These examples are known as the "support vectors", and it is well-known that the weight vector $\boldsymbol{w}^\star$ can be written as a (non-negative) linear combination of the $y_i \boldsymbol{x}_i$ corresponding to support vectors. More precisely, the dual form of the solution expresses the weight vector $\boldsymbol{w}^\star = \sum_{i=1}^n \alpha_i^\star y_i \boldsymbol{x}_i$ in terms of dual variables $\alpha_1^\star, \ldots, \alpha_n^\star \geq 0$. This constitutes a concise representation of the solution—just the list of non-zero dual variables $\alpha_i^\star$ and corresponding data points. This remarkable property of the SVM is particularly important in its "kernelized" extension (Boser et al., 1992; Schölkopf and Smola, 2002), where the dimension d may be very large (or, in fact, infinite) but inner products can be computed efficiently.

The number of support vectors, if sufficiently small, has interesting consequences for the generalization error of the hard-margin SVM solution. Techniques based on leave-one-out analysis and sample compression (Vapnik, 1995; Graepel et al., 2005; Germain et al., 2011) bound the generalization error as a linear function of the fraction of support vectors and have no explicit dependence on the dimension d. In particular, if the number of support vectors can be shown to be $o(n)$ with high probability, these bounds imply "good generalization" of the SVM solution in the sense that the generalization error of the SVM is upper-bounded

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

[1] We only consider homogeneous linear classifiers in this paper and hence have omitted the bias term. The equivalent, but more standard, form of this problem is presented as Eq. (2) in Section 2.1.

by a quantity that tends to zero as $n \to \infty$. Moreover, this sparsity in support vectors can be demonstrated in sufficiently low-dimensional settings using asymptotic arguments (Dietrich et al., 1999; Buhot and Gordon, 2001; Malzahn and Opper, 2005). However, the story is starkly different in the modern high-dimensional (also called *overparameterized*) regime; in fact, quite the opposite can happen. Recent work comparing classification and regression tasks under the high-dimensional linear model (Muthukumar et al., 2020a) showed that under sufficient "effective overparameterization", e.g., $d \sim n \log n$ under isotropic Gaussian design, *every training example is a support vector with high probability.* That is, the fraction of support vectors is exactly 1 with high probability. This establishes a remarkable link between the SVM and solutions that interpolate training data, allowing an entirely different set of recently developed techniques that analyze interpolating solutions in regression tasks (Belkin et al., 2019; Bartlett et al., 2020; Hastie et al., 2019; Mei and Montanari, 2019; Mitra, 2019; Muthukumar et al., 2020b) to be applied to the SVM. Using this equivalence, Muthukumar et al. (2020a) showed the existence of intermediate levels of overparameterization in which all training examples are support vectors with high probability, but the ensuing SVM solution still generalizes well. This characterization was derived for a specific overparameterized ensemble inspired by spiked covariance models (Wang and Fan, 2017; Mahdaviyeh and Naulet, 2019). More importantly, the level of overparameterization considered there was only sufficiently, not necessarily, high enough for support vector proliferation.

In this paper, we establish necessary and sufficient conditions for the phenomenon of support vector proliferation to occur with high probability for a range of high-dimensional linear ensembles, including sub-Gaussian and Haar design of the covariate matrix. In other words, for sufficiently high *effective overparameterization* (measured through quantities that are related to effective ranks of the covariance matrix as identified by Bartlett et al. (2020)), we show that all training examples are support vectors with high probability. We also provide a weak converse: in the absence of a certain level of overparameterization, at least one training example is not a support vector with constant probability.

Related work

The number of support vectors has been previously studied in several contexts on account of the aforementioned connection to generalization error both in classical regimes using sample compression bounds (Vapnik, 1995; Graepel et al., 2005; Germain et al., 2011), and the modern high-dimensional regime (Muthukumar et al., 2020a; Chatterji and Long, 2020). Several works investigate the thermodynamic limit where both the dimension of the input data d and the number of training data n both tend to infinity at a fixed ratio $\delta = n/d$ (e.g., Dietrich et al., 1999; Buhot and Gordon, 2001; Malzahn and Opper, 2005; Liu, 2019). One particular result of note is that of Buhot and Gordon (2001), who consider a linearly[2] separable setting where the training data inputs are drawn iid from a d-dimensional *isotropic* normal distribution. They find that the typical fraction of training examples that are support vectors approaches the following (in the limit as both $n, d \to \infty$):

$$\begin{cases} \frac{0.952}{\delta} & \text{for } \delta \gg 1, \\ 1 - \sqrt{\frac{2\delta}{\pi}} \exp\left(-\frac{1}{2\delta}\right) & \text{for } \delta \ll 1. \end{cases}$$

In the classical regime, where $n \gg d$ (i.e., $\delta \gg 1$), a combination of this asymptotic estimate with sample compression arguments yields generalization error bounds of order $O(1/\delta) = O(d/n)$, which tend to zero as $\delta \to \infty$. However, in the high-dimensional regime, where $d \gg n$ (i.e., $\delta \ll 1$), the fraction of examples that are support vectors quickly approaches 1 as $\delta \to 0$. In these cases, the generalization error bounds based on support vectors no longer provide non-trivial guarantees.

Muthukumar et al. (2020a) recently provided a non-asymptotic result for this isotropic case considered above. They found that if d grows somewhat faster than n (specifically, $d \sim n \log n$), then the fraction of examples that are support vectors is 1 with very high probability. They also showed that the fraction of support vectors obtained by the hard-margin linear SVM can tend to 1 in anisotropic settings *if* the setting is sufficiently high-dimensional; this is captured by notions of effective rank of the covariance matrix of the linear featurizations (Bartlett et al., 2020). Our results greatly sharpen the *sufficient* conditions provided there; see Section 3 for a detailed comparison, and in particular, Section 3.4 for additional discussion of implications for generalization error bounds.

Chatterji and Long (2020) also recently showed that the SVM can generalize well in overparameterized regimes. In their work, the data are generated by a linear model inspired by Fisher's linear discriminant analysis, and establish their results under the assumption of sufficiently high separation between the means of the two classes. Their results are based on a direct analysis of the SVM, but do not make any claims

[2]We note that the main interest of Buhot and Gordon is in SVMs with non-linear feature maps; we quote one of their results specialized to the linear setting.

about the number of support vectors.

The number of support vectors has also been studied in non-separable but low-dimensional settings, using suitable variants of the SVM optimization problem. These variants include the *soft-margin SVM* (Cortes and Vapnik, 1995) and the *ν-SVM* (Schölkopf et al., 2000). In both of these, the hard-margin constraint is relaxed and support vectors include training examples that are exactly on the margin as well as *margin violations*. The soft-margin SVM doees this by introducing slack variables in the margin constraints on examples, and uses a hyper-parameter to control the trade-off between the margin maximization objective and the sum of constraint violations. The ν-SVM provides somewhat more direct control on the number of support vectors: the hyper-parameter ν is an upper-bound on the fraction of margin violations and a lower-bound on the fraction of all support vector examples. First, for a suitable choice of the hyper-parameter, the fraction of examples that are support vectors in the soft-margin SVM can be related to the Bayes error rate when certain kernel functions are used (Steinwart, 2003; Bartlett and Tewari, 2007). Indeed, this fact has motivated algorithmic developments for sparsifying the SVM solution (e.g., Burges, 1996; Downs et al., 2001; Keerthi et al., 2006). Second, under some general conditions on the data distribution, it is also shown for the ν-SVM (Schölkopf et al., 2000, Proposition 5) that as $n \to \infty$ for a fixed dimension d, all support vectors are of the margin violation category. These results for non-separable but low-dimensional settings are not directly comparable to ours, which hold in the high-dimensional (therefore, typically separable) regime. Notably, our results on the support vector proliferation do not require the presence of label noise—i.e., the Bayes error rate can be zero and still, every example may be a support vector.

In addition to the aforementioned sample compression bounds that explicitly use the number of support vectors, there is a distinct line of work on generalization error of SVMs based on the margin γ achieved on the training examples (Bartlett and Shawe-Taylor, 1999; Zhang, 2002; Bartlett and Mendelson, 2002; McAllester, 2003; Grønlund et al., 2020). However, in the settings we consider, these generalization error bounds are never smaller than a universal constant (e.g., $1/\sqrt{2}$), as pointed out by Muthukumar et al. (2020a, Section 6) and expanded upon in Section 3.4. It is worth mentioning that the margin-based bounds, as well as the bounds based on the number of support vectors, make no (or very few) assumptions about the distribution of the training examples. The distribution-free quality makes the bounds widely applicable, but it also limits their ability to capture certain generalization phenomena, such as those from (Muthukumar et al., 2020a; Chatterji and Long, 2020).

Finally, our work bears some resemblance to the early work of Cover (1965) on linear classification. There, the concern is the number of independent features necessary and sufficient for a data set (with fixed, non-random labels) to become linear separable. Linear separability just requires the existence of $\boldsymbol{w} \in \mathbb{R}^d$ such that $\mathrm{margin}_i(\boldsymbol{w}) > 0$ for all $i = 1, \ldots, n$, but these margin values could vary across examples. In contrast, our work considers necessary and sufficient conditions under which the margins achieved are all the same maximum (positive) value.

2 SETTING

In this section, we introduce notation for the SVM problem, and describe the probabilistic models of the training data under which we conduct our analysis.

2.1 SVM optimization problem

Our analysis considers the standard setting for homogeneous binary linear classification with SVMs. In this setting, one has n training examples $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n) \in \mathbb{R}^d \times \{-1, +1\}$. A homogeneous linear classifier is specified by a weight vector $\boldsymbol{w} \in \mathbb{R}^d$, so that the prediction of this classifier on $\boldsymbol{x} \in \mathbb{R}^d$ is given by the sign of $\boldsymbol{x}^\mathsf{T}\boldsymbol{w}$. The ambiguity of the sign when $\boldsymbol{x}^\mathsf{T}\boldsymbol{w} = 0$ is not important in our analysis.

The SVM optimization problem from Eq. (1) is more commonly written as

$$\begin{aligned} \min_{\boldsymbol{w} \in \mathbb{R}^d} \quad & \frac{1}{2}\|\boldsymbol{w}\|_2^2 \\ \text{subj. to} \quad & y_i \boldsymbol{x}_i^\mathsf{T} \boldsymbol{w} \geq 1 \quad \text{for all } i = 1, \ldots, n. \end{aligned} \tag{2}$$

The well-known Lagrangian dual of Eq. (2) can be written entirely in terms of the vector of labels $\boldsymbol{y} := (y_1, \ldots, y_n) \in \mathbb{R}^n$ and the $n \times n$ Gram (or kernel) matrix $\boldsymbol{K}$ corresponding to $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, i.e., $K_{i,j} := \boldsymbol{x}_i^\mathsf{T}\boldsymbol{x}_j$ for all $1 \leq i, j \leq n$:

$$\begin{aligned} \max_{\boldsymbol{\alpha} \in \mathbb{R}^n} \quad & \sum_{i=1}^n \alpha_i - \frac{1}{2}\boldsymbol{\alpha}^\mathsf{T} \operatorname{diag}(\boldsymbol{y})^\mathsf{T} \boldsymbol{K} \operatorname{diag}(\boldsymbol{y}) \boldsymbol{\alpha} \\ \text{subj. to} \quad & \alpha_i \geq 0 \quad \text{for all } i = 1, \ldots, n. \end{aligned} \tag{3}$$

Above, we use $\operatorname{diag}(\cdot)$ to denote the diagonal matrix with diagonal entries taken from the vector-valued argument. An optimal solution $\boldsymbol{\alpha}^\star$ to the dual problem in Eq. (3) corresponds to an optimal primal variable $\boldsymbol{w}^\star$ for the problem in Eq. (2) via the relation $\boldsymbol{w}^\star = \sum_{i=1}^n \alpha_i^\star y_i \boldsymbol{x}_i$. The support vectors are precisely

the examples $(\boldsymbol{x}_i, y_i)$ for which the corresponding $\alpha_i^\star$ is positive, a consequence of complementary slackness.

It will be notationally convenient to change the optimization variable from $\boldsymbol{\alpha}$ to $\boldsymbol{\beta} \in \mathbb{R}^n$ with $\beta_i = y_i \alpha_i$ for all $i = 1, \ldots, n$. In terms of $\boldsymbol{\beta}$, the SVM dual problem from Eq. (3) becomes

$$\begin{aligned} \max_{\boldsymbol{\beta} \in \mathbb{R}^n} \quad & \boldsymbol{y}^\mathsf{T} \boldsymbol{\beta} - \frac{1}{2} \boldsymbol{\beta}^\mathsf{T} \boldsymbol{K} \boldsymbol{\beta} \\ \text{subj. to} \quad & y_i \beta_i \geq 0 \quad \text{for all } i = 1, \ldots, n. \end{aligned} \tag{4}$$

An optimal solution $\boldsymbol{\beta}^\star$ to this problem corresponds to an optimal primal variable $\boldsymbol{w}^\star$ via the relation $\boldsymbol{w}^\star = \sum_{i=1}^n \beta_i^\star \boldsymbol{x}_i$, and the support vectors are precisely the examples $(\boldsymbol{x}_i, y_i)$ for which $\beta_i^\star$ is non-zero.

Note that if it were not for the n constraints, the solutions to optimization problem would be characterized by the linear equation $\boldsymbol{K}\boldsymbol{\beta} = \boldsymbol{y}$. We refer to the version of the optimization problem in Eq. (4) without the n constraints as the *ridgeless regression problem*. Solutions to this problem have been extensively studied in recent years (e.g., Liang and Rakhlin, 2020; Bartlett et al., 2020; Muthukumar et al., 2020b; Belkin et al., 2019; Hastie et al., 2019; Mahdaviyeh and Naulet, 2019). If a vector $\boldsymbol{\beta} \in \mathbb{R}^n$ satisfies both $\boldsymbol{K}\boldsymbol{\beta} = \boldsymbol{y}$ as well as the n constraints $y_i \beta_i \geq 0$ for all $i = 1, \ldots, n$, then $\boldsymbol{\beta}$ is necessarily an optimal solution to the SVM dual problem from Eq. (4).

2.2 Data model

We analyze the SVM under the following probabilistic model of the training examples.

Feature model. The $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are random vectors in $\mathbb{R}^d$ satisfying

$$\boldsymbol{x}_i := \operatorname{diag}(\boldsymbol{\lambda})^{1/2} \boldsymbol{z}_i, \quad \text{for all } i = 1, \ldots, n,$$

The positive vector $\boldsymbol{\lambda} \in \mathbb{R}^d_{++}$ parameterizes the model. The random vectors, collected in the $n \times d$ random matrix $\boldsymbol{Z} := [\boldsymbol{z}_1 | \cdots | \boldsymbol{z}_n]^\mathsf{T} = (z_{i,j})_{1 \leq i \leq n; 1 \leq j \leq d}$, satisfy one of the following distributional assumptions.

1. *Independent features:* $\boldsymbol{Z}$ has independent entries such that each $z_{i,j}$ is mean-zero, unit variance, and sub-Gaussian with parameter $v > 0$ (i.e., $\mathbb{E}(z_{i,j}) = 0$, $\mathbb{E}(z_{i,j}^2) = 1$, and $\mathbb{E}(e^{t z_{i,j}}) \leq e^{v t^2/2}$ for all $t \in \mathbb{R}$).

2. *Haar features:* $\boldsymbol{Z}$ is taken to be the first n rows of a uniformly random $d \times d$ orthogonal matrix (with the Haar measure), and then scaled by $\sqrt{d}$. The scaling is immaterial to our results, but it makes the analysis comparable to that for the independent features case.

Label model. Conditional on $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, the $y_1, \ldots, y_n$ are independent $\{-1, +1\}$-valued random variables such that the conditional distribution of y_i depends only on $\boldsymbol{x}_i$ for each $i = 1, \ldots, n$. Formally:

$$y_i \perp\!\!\!\perp (\boldsymbol{x}_1, y_1, \ldots, \boldsymbol{x}_{i-1}, y_{i-1}, \boldsymbol{x}_{i+1}, y_{i+1}, \ldots, \boldsymbol{x}_n, y_n) \mid \boldsymbol{x}_i.$$

Remarks. All of our results will assume $d \geq n$. The non-singularity of the kernel matrix $\boldsymbol{K} = \boldsymbol{Z} \operatorname{diag}(\boldsymbol{\lambda}) \boldsymbol{Z}^\mathsf{T}$ will be important for our analysis. In the case of Haar features, setting $d \geq n$ ensures that the matrix $\boldsymbol{Z}$ always has rank n, and hence the kernel matrix $\boldsymbol{K} = \boldsymbol{Z} \operatorname{diag}(\boldsymbol{\lambda}) \boldsymbol{Z}^\mathsf{T}$ is always non-singular. In the case of independent features, if the distributions of the $z_{i,j}$ are continuous, then $\boldsymbol{Z}$ has rank n almost surely, and hence again $\boldsymbol{K}$ is non-singular almost surely. Our results only require the $z_{i,j}$ to be sub-Gaussian and need not have continuous distributions. For instance, if the $z_{i,j}$ are Rademacher (uniform on $\{-1, +1\}$), then there is a non-zero probability that $\boldsymbol{Z}$ is rank-deficient—however, we will see that this probability is negligible.

Our label model is very general and allows for a variety of settings, including the following.

1. *Generalized linear models (GLMs):* $\Pr(y_i = 1 \mid \boldsymbol{x}_i) = g(\boldsymbol{x}_i^\mathsf{T} \boldsymbol{w})$ for some $\boldsymbol{w} \in \mathbb{R}^d$ and some function $g \colon \mathbb{R} \to [0, 1]$. Examples include *logistic regression*, where $g(t) = 1/(1 + e^{-t})$; *probit regression*, where $g(t) = \Phi(t)$ and Φ is the cumulative distribution function of the standard Gaussian distribution; and *one-bit compressive sensing* (Boufounos and Baraniuk, 2008), where $g(t) = \mathbb{1}_{\{t > 0\}}$.

2. *Multi-index models:* $\Pr(y_i = 1 \mid \boldsymbol{x}_i) = h(\boldsymbol{W}\boldsymbol{x}_i)$ for some $k \in \mathbb{N}$, $\boldsymbol{W} \in \mathbb{R}^{k \times d}$, and $h \colon \mathbb{R}^k \to [0, 1]$. The case $k = 1$ corresponds to GLMs. Examples with $k \geq 2$ include the intersections of half-spaces models and certain neural networks (Baum, 1990; Klivans et al., 2004; Klivans and Servedio, 2008).

3. *Fixed labels:* $y_i \in \{-1, +1\}$ are fixed (non-random) values. This can be regarded as a null model where the feature vectors have no statistical relationship to the labels. This null model was, e.g., considered by Cover (1965).

Our results in Theorem 1 and Theorem 2 consider, respectively, the independent features and Haar features, but both allowing for general label models. Our weak converse result in Theorem 3 is established in the special case where the $z_{i,j}$ are iid standard Gaussian random variables (a special case of independent features), and where the labels are fixed.

2.3 Additional notation

Let $[n] := \{1, \dots, n\}$ for any natural number n. Let $\mathbb{R}_{++} := \{x \in \mathbb{R} \colon x > 0\}$ denote the positive real numbers. For a vector $\boldsymbol{v} \in \mathbb{R}^n$, we let $\boldsymbol{v}_{\setminus i} \in \mathbb{R}^{n-1}$ denote the vector obtained from $\boldsymbol{v}$ by omitting the i^{th} coordinate. For a matrix $\boldsymbol{M} \in \mathbb{R}^{n \times d}$, we let $\boldsymbol{M}_{\setminus i} \in \mathbb{R}^{(n-1)\times d}$ denote the matrix obtained from $\boldsymbol{M}$ by omitting the i^{th} row. Sometimes, for a square matrix $\boldsymbol{M} \in \mathbb{R}^{n \times n}$, we will also use $\boldsymbol{M}_{\setminus i} \in \mathbb{R}^{(n-1)\times(n-1)}$ to denote the matrix obtained from $\boldsymbol{M}$ by removing the i^{th} row and column. We let $\boldsymbol{e}_i$ denote the i^{th} coordinate vector in $\mathbb{R}^n$. For a vector $\boldsymbol{v} \in \mathbb{R}^d$, we denote its p-norm by $\|\boldsymbol{v}\|_p = (\sum_{i=1}^d |v_i|^p)^{1/p}$. For a matrix $\boldsymbol{M} \in \mathbb{R}^{d \times d}$, we denote its $2 \to 2$ operator norm (i.e., largest singular value) by $\|\boldsymbol{M}\|_{\text{op}} = \sup_{\boldsymbol{v} \in \mathbb{R}^d : \|\boldsymbol{v}\|_2 \le 1} \|\boldsymbol{M}\boldsymbol{v}\|_2$. Let $S^{d-1} := \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_2 = 1\}$ denote the unit sphere in $\mathbb{R}^d$. If $\boldsymbol{M}$ is a symmetric matrix, $\lambda_{\min}(\boldsymbol{M})$ denotes the smallest eigenvalue of $\boldsymbol{M}$. Finally, we will use (C, c, c_1, c_2) to denote universal constants that do not depend, explicitly or implicitly, on the dimension d, the number of training examples n, or properties of the data distribution.

3 MAIN RESULTS

Our primary interest is in the probability that every training example is a support vector under the data model from Section 2.2. We give sufficient conditions on certain effective dimensions for this probability to tend to one as $n \to \infty$. We complement these results with a partial weak converse. Finally, we present a key deterministic result that is used in the proofs of the aforementioned results. All proofs are given in Appendix A.

We define the following effective dimensions in terms of the data model parameter $\boldsymbol{\lambda}$:

$$d_2 := \frac{\|\boldsymbol{\lambda}\|_1^2}{\|\boldsymbol{\lambda}\|_2^2} \qquad \text{and} \qquad d_\infty := \frac{\|\boldsymbol{\lambda}\|_1}{\|\boldsymbol{\lambda}\|_\infty}.$$

Observe that $d \ge d_2 \ge d_\infty$, and that if $\lambda_j = 1$ for all $j = 1, \dots, d$ (i.e., the isotropic setting), then $d = d_2 = d_\infty$. We note that d_2 and d_∞ are, respectively, the same as the effective ranks $r_0(\operatorname{diag}(\boldsymbol{\lambda}))$ and $R_0(\operatorname{diag}(\boldsymbol{\lambda}))$ studied by Bartlett et al. (2020). They arise naturally from the tail behavior of certain linear combinations of χ^2-random variables (see, e.g., Laurent and Massart, 2000).

3.1 Sufficient conditions

Our first main result provides sufficient conditions on the effective dimensions d_2 and d_∞ in the independent features setting so that, with probability tending to one, every training example is a support vector.

Theorem 1. *There are universal constants $C > 0$ and $c > 0$ such that the following holds. If the training data $(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_n, y_n)$ follow the model from Section 2.2 with independent features, subgaussian parameter $v > 0$, and model parameter $\boldsymbol{\lambda} \in \mathbb{R}^d_{++}$, then the probability that every training example is a support vector is at least*

$$1 - \exp\left(-c \cdot \min\left\{\frac{d_2}{v^2}, \frac{d_\infty}{v}\right\} + Cn\right) - \exp\left(-c \cdot \frac{d_\infty}{vn} + C \log n\right).$$

Observe that the probability from Theorem 1 is close to 1 when

$$d_2 \gg v^2 n \qquad \text{and} \qquad d_\infty \gg vn \log n.$$

We can compare this condition to that from the prior work of Muthukumar et al. (2020a) in our setting with independent Gaussian features ($v = 1$). In the anisotropic setting (i.e., general $\boldsymbol{\lambda}$), the prior result's condition for every training example to be a support vector with high probability is $d_2 \gg n^2 \log n$ and $d_\infty \gg n^{3/2} \log n$. In the isotropic setting (i.e., all $\lambda_j = 1$), assuming the labels are fixed (i.e., non-random), the prior result's condition is $d \gg n \log n$. Theorem 1 is an improvement in the anisotropic case, and it matches this prior result in the isotropic case.[3]

Our second main result provides an analogue of Theorem 1 for the case of Haar features (where neither training examples nor features are statistically independent).

Theorem 2. *There are universal constants $C > 0$ and $c > 0$ such that the following holds. If the training data $(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_n, y_n)$ follow the model from Section 2.2 with Haar features and model parameter $\boldsymbol{\lambda} \in \mathbb{R}^d_{++}$, then the probability that every training example is a support vector is at least*

$$1 - \exp\left(-c \cdot d_\infty + Cn\right) - \exp\left(-c \cdot \frac{d - n + 1}{d} \cdot \frac{d_\infty}{n} + C \log n\right).$$

3.2 Weak converse

Our final main result gives a weak converse to Theorem 1 in the case where the features are iid standard Gaussians and the labels are fixed.

[3] We remark that the result of Muthukumar et al. (2020a) for the anisotropic case, in fact, holds for all (fixed) label vectors $\boldsymbol{y} \in \{-1, +1\}^n$ simultaneously. However, their proof does not readily give a tighter condition when only a single (random) label vector is considered. Our proof technique side-steps this issue by showing that it is sufficient to consider the scaling of quantities that do not depend on the value of the label vector.

Theorem 3. *Let the training data $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ follow the model from Section 2.2 with $\boldsymbol{\lambda} = (1, \ldots, 1)$, $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ being iid standard Gaussian random vectors in $\mathbb{R}^d$, and $y_1, \ldots, y_n \in \{\pm 1\}$ being arbitrary but fixed (i.e., non-random) values. For any $d \geq n$, the probability that at least one training example is not a support vector is at least*

$$\Phi\left(-\sqrt{\frac{d-n+4+2\sqrt{d-n+2}}{n-1}}\right) \cdot \left(1 - \frac{1}{e}\right),$$

where Φ is the cumulative distribution function of the standard Gaussian distribution.

Observe that the probability bound from Theorem 3 is at least a positive constant (independent of d and n) whenever the dimension d (regarded as a function of n) is $O(n)$. This means that the dimension d must be super-linear in n in order for the "success" probability of Theorem 1 to tend to one with n.

Theorem 3 applies to the case where the features vectors are isotropic. In Appendix B, we give a version of the result that applies to certain anisotropic settings, again in the case of independent Gaussian features and fixed labels. The theorem puts restrictions on the tail behavior of $\boldsymbol{\lambda}$. These restrictions are related to the effective ranks studied by Bartlett et al. (2020). The proof is similar to that of Theorem 3, but also relies on a technical result from (Bartlett et al., 2020).

Except when the "success" probability is required to be $\geq 1 - 1/n^c$ for constant $c > 0$, there is a $\log(n)$ gap between the sufficient condition from Theorem 1 and the necessary condition from Theorem 3. Our approach to the proof of Theorem 3 does not appear to be able to close this gap. We discuss this issue further in Appendix C, and leave its resolution to future work.

3.3 Deterministic equivalences

The crux of all of the above results lies in the following key lemma, which characterizes equivalent conditions for every training example to be a support vector.

Lemma 1. *Suppose $\boldsymbol{Z} := [\boldsymbol{z}_1 | \cdots | \boldsymbol{z}_n]^\top \in \mathbb{R}^{n \times d}$ and $\boldsymbol{\lambda} \in \mathbb{R}^d_{++}$ are such that $\boldsymbol{Z} \operatorname{diag}(\boldsymbol{\lambda}) \boldsymbol{Z}^\top$ and $\boldsymbol{Z}_{\setminus i} \operatorname{diag}(\boldsymbol{\lambda}) \boldsymbol{Z}_{\setminus i}^\top$ for all $i = 1, \ldots, n$ are non-singular. Let the training data $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n) \in \mathbb{R}^d \times \{-1, +1\}$ satisfy $\boldsymbol{x}_i = \operatorname{diag}(\boldsymbol{\lambda})^{1/2} \boldsymbol{z}_i$ for each $i = 1, \ldots, n$. Then the following are equivalent:*

1. *Every training example is a support vector.*
2. *The vector $\boldsymbol{\beta} := \boldsymbol{K}^{-1} \boldsymbol{y}$ satisfies $y_i \beta_i > 0$ for all $i = 1, \ldots, n$.*
3. *$y_i \boldsymbol{y}_{\setminus i}^\top \left(\boldsymbol{Z}_{\setminus i} \operatorname{diag}(\boldsymbol{\lambda}) \boldsymbol{Z}_{\setminus i}^\top\right)^{-1} \boldsymbol{Z}_{\setminus i} \operatorname{diag}(\boldsymbol{\lambda}) \boldsymbol{z}_i < 1$ for all $i = 1, \ldots, n$.*

The above lemma is a deterministic result—it does not reference a particular statistical model for the data—and hence the equivalences are given under non-singularity conditions. We note that the non-singularity conditions are readily satisfied under the data model from Section 2.2 (with high probability, in the case of independent features, or deterministically, in the case of Haar features).

The equivalences of the first two items in this lemma connect the solutions to the SVM optimization problem and the ridgeless regression problem more tightly than was done in the prior work of Muthukumar et al. (2020a), who only proved one direction of the equivalence between the first two items. The proofs of our main results critically use the third item in the above equivalence.

3.4 Implications for generalization

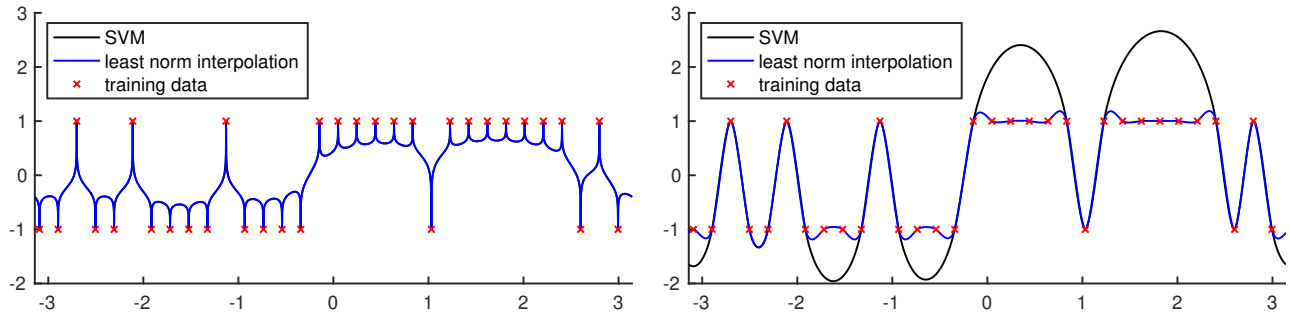
In Theorem 1 and Theorem 2, we identified high-dimensional regimes in which the SVM solution exactly corresponds to the least norm (linear) interpolation of training data with high probability. We observe in Figure 1 that certain deterministic featurizations (which bear some resemblance to the Haar features of Theorem 2, and have been independently analyzed in the interpolating regime for regression problems (Belkin et al., 2019; Muthukumar et al., 2020b)) also empirically exhibit similar support vector proliferation when the effective overparameterization is sufficiently high.

The regimes considered in our results go beyond the common high-dimensional asymptotic where d and n grow proportionally to each other (i.e., $n/d \to \delta$ as $n, d \to \infty$). One may wonder, then, whether these regimes are too high dimensional for the SVM to generalize well. As mentioned in Section 1, the classical generalization error bounds for the SVM are based on the number of support vectors or the worst-case margin achieved on the training examples. Recall that these upper bounds are, respectively, roughly of the form[4]

$$\frac{\#\text{ support vectors}}{n} \quad \text{and} \quad \frac{\|\boldsymbol{w}^\star\|_2^2}{n} \cdot \frac{\mathbb{E}[\operatorname{tr}(\boldsymbol{K})]}{n}.$$

Here, $\boldsymbol{w}^\star$ is the solution to the SVM primal problem

[4]Some bounds are given as the square-roots of the expressions we show, but whether or not the square-root is used will not make a difference in our case. We also omit constants (which are typically larger than 1), polylogarithmic factors in n, and terms related to the confidence level for the bound.



(a) $\eta_i = 1/i$ (b) $\eta_i = 1/i^3$

Figure 1: Plots of linear functions on top of trigonometric features of a scalar input variable that parameterizes the horizontal axis. (These plots originally appeared in (Muthukumar et al., 2020a).) The two linear functions are those given by the solution to the SVM optimization problem and the ridgeless regression problem (i.e., the least norm interpolation), based on 32 training data shown as $\times$'s in the plot. The features are obtained via the mapping $t \mapsto (1, \sqrt{\eta_1}\cos(1\cdot t), \sqrt{\eta_1}\sin(1\cdot t), \ldots, \sqrt{\eta_k}\cos(k\cdot t), \sqrt{\eta_k}\sin(k\cdot t)) \in \mathbb{R}^{2k+1}$ where $k = 2^{14}$. In (a), the SVM and least norm interpolation coincide exactly (so all 32 examples are support vectors); in (b), the functions are noticeably distinct (and only 18 out of 32 examples are support vectors). In each case, we computed analogues of d_2 and d_∞ based on the eigenvalues of the Gram matrix. In (a), they are 108.386 and 21.5626; in (b), they are 3.21378 and 2.20198.

in Eq. (2). Unfortunately, these bounds are not informative for the high-dimensional regimes in which all training points become support vectors. As soon as d_2 and d_∞, respectively, grow beyond n and $n \log n$, then both bounds above become trivial with probability tending to one. This is immediately apparent for the first bound, as a consequence of Theorem 1. For the second bound, an inspection of the proof of Theorem 1 shows that in an event where every training example is a support vector (with the same probability as given in Theorem 1), we have

$$\|\boldsymbol{w}^\star\|_2^2 = \boldsymbol{y}^\top \boldsymbol{K}^{-1}\boldsymbol{y} \geq \frac{n}{\|\boldsymbol{K}\|_{\mathrm{op}}} \geq \frac{n}{2\|\boldsymbol{\lambda}\|_1}.$$

Since $\mathbb{E}[\mathrm{tr}\,(\boldsymbol{K})]/n = \|\boldsymbol{\lambda}\|_1$, the second bound is at least $1/2$ in this event. We also remark that even more sophisticated generalization bounds using the distribution of the margin on training examples (e.g., Gao and Zhou, 2013) do not help in this high-dimensional regime. This is because when all training examples become support vectors, the normalized margin of every training point becomes exactly the worst-case margin, which is $1/\|\boldsymbol{w}^\star\|_2$.

However, recent analyses show that the SVM can generalize well even when all training points become support vectors. In particular, the recent work of Muthukumar et al. (2020a) provided positive implications for the SVM by analyzing the classification test error of the least norm interpolation. In particular, they considered a special anisotropic Gaussian ensemble inspired by spiked covariance models, parameterized by positive constants $p > 1$ and $0 < (q, r) < 1$; here, $d = n^p$ and (q, r) parameterize the eigenvalues of the feature covariance matrix and the sparsity of the unknown signal respectively. See (Muthukumar et al., 2020a, Section 3.4) for further details. It suffices for our purposes to note that the main result of Muthukumar et al. (2020a, Theorem 2) showed that the following *rate region* of (p, q, r) is necessary and sufficient for the least norm interpolation of training data to generalize well, in the sense that the classification test error goes to 0 as $n \to \infty$:

$$0 \leq q < 1 - r + \frac{p-1}{2}. \tag{5}$$

It is easy to verify that Theorem 1 directly implies good generalization of the SVM for this entire rate region. First, for $q \geq 1 - r$, it holds that

$$d_2 \asymp n^{2p - \max\{2p-2q-r, p\}}$$
$$d_\infty \asymp n^{q+r},$$

and since we have assumed $p > 1$, the conditions of Theorem 1, i.e., $d_2 \gg n$, $d_\infty \gg n \log n$, would hold if and only if $q > 1 - r$. On the other hand, the usual margin-based bounds would show good generalization of the SVM *if* $0 \leq q < (1 - r)$. Putting these together, the SVM generalizes well for the entire rate region in Equation (5).

Further, the improvement of this implication over the partial implications for the SVM that were provided in Muthukumar et al. (2020a) is clear. In particular, (Muthukumar et al., 2020a, Corollary 1) required $p > 2$, i.e. $d \gg n^2$, and showed that the SVM will then generalize well if $(3/2 - r) < q < (1 - r) + (p - 1)/2$.

Thus, the rate region implied by this work was

$$\{0 \le q < (1-r)\} \cup \left\{ \left(\frac{3}{2} - r\right) < q < (1-r) + \frac{(p-1)}{2} \right\},$$

which has a non-trivial gap compared to Eq. (5). In summary, our results imply an expansion over the rate region predicted by classical generalization bounds based on either the number of support vectors or the margin.

4 PROOF OF MAIN LEMMA

This section gives the proof of the main technical lemma (Lemma 1). (The proofs of the main results are given in Appendix A.)

Throughout, we use the shorthand notations $\boldsymbol{\Lambda} := \operatorname{diag}(\boldsymbol{\lambda})$ and $\boldsymbol{K}_{\setminus i} := \boldsymbol{Z}_{\setminus i}\boldsymbol{\Lambda}\boldsymbol{Z}_{\setminus i}^\mathsf{T}$ for each $i = 1, \dots, n$. Note that $\boldsymbol{K}_{\setminus i}$ is the same as $\boldsymbol{K} = \boldsymbol{Z}\operatorname{diag}(\boldsymbol{\lambda})\boldsymbol{Z}^\mathsf{T}$ except omitting both the i^{th} row *and* the i^{th} column.

We now give the proof of Lemma 1. Recall that we assume $\boldsymbol{K}$ and $\boldsymbol{K}_{\setminus i}$ for all $i = 1, \dots, n$ are non-singular. We first show that all training examples are support vectors if and only if the candidate solution $\boldsymbol{\beta} = \boldsymbol{K}^{-1}\boldsymbol{y}$ satisfies

$$y_i\beta_i > 0 \quad \text{for all } i = 1, \dots, n. \tag{6}$$

($\Longleftarrow$) Assume $y_i\beta_i > 0$ for all $i \in [n]$. Recall that $\boldsymbol{\beta} = \boldsymbol{K}^{-1}\boldsymbol{y}$ is the unique optimal solution to the ridgeless regression problem (i.e., the problem in Eq. (4) without the n constraints). Since Eq. (6) holds, then $\boldsymbol{\beta}$ is dual-feasible as well, and so it is the unique optimal solution to the dual program, i.e., $\boldsymbol{\beta}^\star = \boldsymbol{\beta}$. Moreover, $y_i\beta_i^\star > 0 \implies \beta_i^\star \neq 0$ for all $i \in [n]$, and so every training example is a support vector.

($\Longrightarrow$) Assume every training example is a support vector, i.e., $\beta_i^\star \neq 0$ for all $i \in [n]$ (so, in particular, $y_i\beta_i^\star > 0$ for all $i \in [n]$). We shall write the solution $\boldsymbol{w}^\star$ to the primal problem from Eq. (2) as a linear combination of $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ in two ways. The first way is in terms of the dual solution $\boldsymbol{\beta}^\star$, i.e., $\boldsymbol{w}^\star = \sum_{i=1}^n \beta_i^\star \boldsymbol{x}_i$, which follows by strong duality. The second way comes via complementary slackness, which implies that $\boldsymbol{w}^\star$ satisfies every constraint in Eq. (2) with equality. In other words, $\boldsymbol{w}^\star$ solves $\min_{\boldsymbol{w}\in\mathbb{R}^d} \frac{1}{2}\|\boldsymbol{w}\|_2^2$ subject to $\boldsymbol{x}_i^\mathsf{T}\boldsymbol{w} = y_i$ for all $i = 1, \dots, n$. Since $\boldsymbol{K}$ is non-singular by assumption, the solution is unique and is given by $\boldsymbol{X}^\mathsf{T}\boldsymbol{K}^{-1}\boldsymbol{y} = \boldsymbol{X}^\mathsf{T}\boldsymbol{\beta} = \sum_{i=1}^n \beta_i\boldsymbol{x}_i$, where $\boldsymbol{X} = [\boldsymbol{x}_1|\cdots|\boldsymbol{x}_n]^\mathsf{T}$. So we have $\boldsymbol{w}^\star = \sum_{i=1}^n \beta_i^\star\boldsymbol{x}_i = \sum_{i=1}^n \beta_i\boldsymbol{x}_i$. The non-singularity of $\boldsymbol{K}$ also implies that $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ are linearly independent, so we must have $\beta_i = \beta_i^\star \neq 0$ for all $i \in [n]$, and thus Eq. (6) holds.

So we have shown that all training examples are support vectors if and only if Eq. (6) holds. It therefore suffices to show that, for each $i = 1, \dots, n$,

$$y_i\beta_i > 0 \quad \Longleftrightarrow \quad y_i\boldsymbol{y}_{\setminus i}^\mathsf{T}\boldsymbol{K}_{\setminus i}^{-1}\boldsymbol{Z}_{\setminus i}\boldsymbol{\Lambda}\boldsymbol{z}_i < 1.$$

By symmetry, we only need to show this implication for $i = 1$.

Observe that $y_1\beta_1 = y_1\boldsymbol{e}_1^\mathsf{T}\boldsymbol{K}^{-1}\boldsymbol{y} = \boldsymbol{e}_1^\mathsf{T}\boldsymbol{K}^{-1}(y_1\boldsymbol{y})$ is the inner product between the first row of $\boldsymbol{K}^{-1}$ and $y_1\boldsymbol{y}$. Therefore, by Cramer's rule, we have

$$y_1\beta_1 = y_1\boldsymbol{e}_1^\mathsf{T}\boldsymbol{K}^{-1}\boldsymbol{y} = \frac{\det(\tilde{\boldsymbol{K}})}{\det(\boldsymbol{K})}$$

where $\tilde{\boldsymbol{K}}$ is the matrix obtained from $\boldsymbol{K}$ by replacing the first row with $y_1\boldsymbol{y}^\mathsf{T}$. Since $\boldsymbol{K}$ is assumed to be invertible, $\boldsymbol{K}$ is positive definite, and so $\det(\boldsymbol{K}) > 0$. Hence, we have $y_1\beta_1 > 0$ *iff* $\det(\tilde{\boldsymbol{K}}) > 0$.

Let us write $\tilde{\boldsymbol{K}}$ as

$$\tilde{\boldsymbol{K}} = \begin{bmatrix} 1 & y_1\boldsymbol{y}_{\setminus 1}^\mathsf{T} \\ \boldsymbol{a} & \boldsymbol{K}_{\setminus 1} \end{bmatrix},$$

where $\boldsymbol{a} := \boldsymbol{Z}_{\setminus 1}\boldsymbol{\Lambda}\boldsymbol{z}_1$ and recall that $\boldsymbol{K}_{\setminus 1}$ denotes the $(n-1)\times(n-1)$ matrix obtained by removing the first row and column from $\boldsymbol{K}$. Note that $\boldsymbol{K}_{\setminus 1}$ is invertible by assumption and hence positive definite. Also, define

$$\boldsymbol{Q} := \begin{bmatrix} 1 & -y_1\boldsymbol{y}_{\setminus 1}^\mathsf{T} \\ \boldsymbol{0} & \boldsymbol{I}_{n-1} \end{bmatrix},$$

where $\boldsymbol{I}_{n-1}$ is the $(n-1)\times(n-1)$ identity matrix. Every diagonal entry of $\boldsymbol{Q}$ is equal to 1, so $\det(\boldsymbol{Q}) = 1$. Hence

$$\begin{aligned} \det(\tilde{\boldsymbol{K}}) &= \det(\tilde{\boldsymbol{K}})\det(\boldsymbol{Q}) \\ &= \det(\tilde{\boldsymbol{K}}\boldsymbol{Q}) \\ &= \det\left(\begin{bmatrix} 1 & \boldsymbol{0}^\mathsf{T} \\ \boldsymbol{a} & \boldsymbol{K}_{\setminus 1} - y_1\boldsymbol{a}\boldsymbol{y}_{\setminus 1}^\mathsf{T} \end{bmatrix}\right) \\ &= \det(\boldsymbol{K}_{\setminus 1} - \boldsymbol{a}\boldsymbol{b}^\mathsf{T}) \end{aligned}$$

where $\boldsymbol{b} := y_1\boldsymbol{y}_{\setminus 1}$. Therefore, $\det(\tilde{\boldsymbol{K}}) > 0$ *iff* $\det(\boldsymbol{K}_{\setminus 1} - \boldsymbol{a}\boldsymbol{b}^\mathsf{T}) > 0$.

By the matrix determinant lemma,

$$\det(\boldsymbol{K}_{\setminus 1} - \boldsymbol{a}\boldsymbol{b}^\mathsf{T}) = \det(\boldsymbol{K}_{\setminus 1})(1 - \boldsymbol{b}^\mathsf{T}\boldsymbol{K}_{\setminus 1}^{-1}\boldsymbol{a}).$$

Since $\boldsymbol{K}_{\setminus 1}$ is positive definite, we have $\det(\boldsymbol{K}_{\setminus 1}) > 0$. Hence, $\det(\boldsymbol{K}_{\setminus 1} - \boldsymbol{a}\boldsymbol{b}^\mathsf{T}) > 0$ *iff* $\boldsymbol{b}^\mathsf{T}\boldsymbol{K}_{\setminus 1}^{-1}\boldsymbol{a} < 1$.

Connecting all of the equivalences and plugging-in for $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{K}_{\setminus 1}$, we have shown that

$$y_1\beta_1 > 0 \quad \Longleftrightarrow \quad y_1\boldsymbol{y}_{\setminus 1}^\mathsf{T}(\boldsymbol{Z}_{\setminus 1}\boldsymbol{\Lambda}\boldsymbol{Z}_{\setminus 1}^\mathsf{T})^{-1}\boldsymbol{Z}_{\setminus 1}\boldsymbol{\Lambda}\boldsymbol{z}_1 < 1,$$

as required. This completes the proof of the lemma. □

Acknowledgements

This project resulted from a collaboration initiated during the "Foundations of Deep Learning" program at the Simons Institute for the Theory of Computing, and we are grateful to the Institute and organizers for their hospitality and support of such collaborative research. We thank Clayton Sanford for his careful reading and comments on this paper. DH acknowledges partial support from NSF awards CCF-1740833 and IIS-1815697, a Sloan Research Fellowship, and a Google Faculty Award. VM acknowledges partial support from a Simons-Berkeley Research Fellowship, support of the ML4Wireless center member companies and NSF grants AST-144078 and ECCS-1343398. JX was supported by a Cheung-Kong Graduate School of Business Fellowship.

References

Bartlett, P. and Shawe-Taylor, J. (1999). Generalization performance of support vector machines and other pattern classifiers. In *Advances in Kernel Methods: Support Vector Learning*. MIT Press, Cambridge, MA.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. (2020). Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*.

Bartlett, P. L. and Mendelson, S. (2002). Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482.

Bartlett, P. L. and Tewari, A. (2007). Sparseness vs estimating conditional probabilities: Some asymptotic results. *Journal of Machine Learning Research*, 8(Apr):775–790.

Baum, E. B. (1990). A polynomial time algorithm that learns two hidden unit nets. *Neural Computation*, 2(4):510–522.

Belkin, M., Hsu, D., and Xu, J. (2019). Two models of double descent for weak features. *arXiv preprint arXiv:1903.07571*.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages 144–152.

Boufounos, P. T. and Baraniuk, R. G. (2008). 1-bit compressive sensing. In *2008 42nd Annual Conference on Information Sciences and Systems*, pages 16–21. IEEE.

Buhot, A. and Gordon, M. B. (2001). Robust learning and generalization with support vector machines. *Journal of Physics A: Mathematical and General*, 34(21):4377.

Burges, C. J. (1996). Simplified support vector decision rules. In *International Conference on Machine Learning*, volume 96, pages 71–77.

Chatterji, N. S. and Long, P. M. (2020). Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *arXiv preprint arXiv:2004.12019*.

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

Cover, T. M. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, 14(3):326–334.

Dietrich, R., Opper, M., and Sompolinsky, H. (1999). Statistical mechanics of support vector networks. *Physical Review Letters*, 82(14):2975.

Downs, T., Gates, K. E., and Masters, A. (2001). Exact simplification of support vector solutions. *Journal of Machine Learning Research*, 2(Dec):293–297.

Gao, W. and Zhou, Z.-H. (2013). On the doubt about margin explanation of boosting. *Artificial Intelligence*, 203:1–18.

Germain, P., Lacoste, A., Laviolette, F., Marchand, M., and Shanian, S. (2011). A PAC-Bayes sample-compression approach to kernel methods. In *ICML*.

Graepel, T., Herbrich, R., and Shawe-Taylor, J. (2005). PAC-Bayesian compression bounds on the prediction error of learning algorithms for classification. *Machine Learning*, 59(1-2):55–76.

Grønlund, A., Kamma, L., and Larsen, K. G. (2020). Near-tight margin-based generalization bounds for support vector machines. *arXiv preprint arXiv:2006.02175*.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. (2019). Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*.

Keerthi, S. S., Chapelle, O., and DeCoste, D. (2006). Building support vector machines with reduced classifier complexity. *Journal of Machine Learning Research*, 7(Jul):1493–1515.

Klivans, A. R., O'Donnell, R., and Servedio, R. A. (2004). Learning intersections and thresholds of halfspaces. *Journal of Computer and System Sciences*, 68(4):808–840.

Klivans, A. R. and Servedio, R. A. (2008). Learning intersections of halfspaces with a margin. *Journal of Computer and System Sciences*, 74(1):35–48.

Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338.

Liang, T. and Rakhlin, A. (2020). Just interpolate: Kernel "ridgeless" regression can generalize. *Annals of Statistics*, 48(3):1329–1347.

Liu, H. (2019). Exact high-dimensional asymptotics for support vector machine. *arXiv preprint arXiv:1905.05125*.

Mahdaviyeh, Y. and Naulet, Z. (2019). Risk of the least squares minimum norm estimator under the spike covariance model. *arXiv preprint arXiv:1912.13421*.

Malzahn, D. and Opper, M. (2005). A statistical physics approach for the analysis of machine learning algorithms on real data. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(11):P11001.

McAllester, D. (2003). Simplified PAC-Bayesian margin bounds. In *Learning Theory and Kernel Machines*, pages 203–215. Springer.

Mei, S. and Montanari, A. (2019). The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*.

Mitra, P. P. (2019). Understanding overfitting peaks in generalization error: Analytical risk curves for ℓ_2 and ℓ_1 penalized interpolation. *arXiv preprint arXiv:1906.03667*.

Muthukumar, V., Narang, A., Subramanian, V., Belkin, M., Hsu, D., and Sahai, A. (2020a). Classification vs regression in overparameterized regimes: Does the loss function matter? *arXiv preprint arXiv:2005.08054*.

Muthukumar, V., Vodrahalli, K., Subramanian, V., and Sahai, A. (2020b). Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 1(1):67–83.

Pisier, G. (1999). *The volume of convex bodies and Banach space geometry*, volume 94. Cambridge University Press.

Rudelson, M. and Vershynin, R. (2013). Hanson-Wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18.

Schölkopf, B. and Smola, A. J. (2002). *Learning with kernels*. MIT Press.

Schölkopf, B., Smola, A. J., Williamson, R. C., and Bartlett, P. L. (2000). New support vector algorithms. *Neural computation*, 12(5):1207–1245.

Steinwart, I. (2003). Sparseness of support vector machines. *Journal of Machine Learning Research*, 4(Nov):1071–1105.

Vapnik, V. N. (1982). *Estimation of dependences based on empirical data*. Springer-Verlag.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

Wang, W. and Fan, J. (2017). Asymptotics of empirical eigenstructure for high dimensional spiked covariance. *Annals of Statistics*, 45(3):1342.

Zhang, T. (2002). Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2(Mar):527–550.