

PAPER

On the proliferation of support vectors in high dimensions^{*}

To cite this article: Daniel Hsu *et al* *J. Stat. Mech.* (2022) 114011

View the [article online](#) for updates and enhancements.

You may also like

- [A new formulation of gradient boosting](#)
Alex Wozniakowski, Jayne Thompson, Mile Gu et al.
- [Fast approximations of spectral lineshapes to enable optimization of a filtered Rayleigh scattering experiment](#)
Gregory J Hunt, Cody R Ground and Robin L Hunt
- [A fast sparse least squares support vector machine hysteresis model for piezoelectric actuator](#)
Xuefei Mao, Haocheng Du, Siwei Sun et al.

PAPER: ML 2022

On the proliferation of support vectors in high dimensions*

Daniel Hsu^{1,2}, Vidya Muthukumar^{3,4,**} and Ji Xu¹

¹ Department of Computer Science, Columbia University, New York, United States of America

² Data Science Institute, Columbia University, New York, United States of America

³ School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, United States of America

⁴ School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, United States of America

E-mail: vmuthukumar8@gatech.edu

Received 15 June 2021

Accepted for publication 4 July 2022

Published 24 November 2022



CrossMark

Online at stacks.iop.org/JSTAT/2022/114011

<https://doi.org/10.1088/1742-5468/ac98a9>

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.

Keywords: learning theory, machine learning, statistical inference

*This article is an updated version of: Hsu D, Muthukumar V and Xu J L 2021 On the proliferation of support vectors in high dimensions *Proc. 24th Int. Conf. Artificial Intelligence and Statistics* vol 130 ed Banerjee A and Fukumizu K pp 91–9.

**Author to whom any correspondence should be addressed.

Dimension 7

Contents

1. Introduction	2
1.1. Related work.....	4
2. Setting	6
2.1. SVM optimization problem	6
2.2. Data model.....	7
2.3. Additional notation.....	8
3. Main results	9
3.1. Sufficient conditions.....	9
3.2. Weak converse.....	10
3.3. Deterministic equivalences	11
3.4. Implications for generalization	11
4. Proofs	13
4.1. Proof of lemma 1	14
4.2. Proof of theorem 1.....	15
4.3. Proof of theorem 2.....	18
4.4. Proof of theorem 3.....	20
Acknowledgments.....	21
Appendix A. Anisotropic version of theorem 3.....	22
Appendix B. Probabilistic inequalities	24
References.....	26

1. Introduction

The support vector machine (SVM) is one of the most well-known and commonly used methods for binary classification in machine learning [14, 42]. 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 $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n) \in \mathbb{R}^d \times \{-1, +1\}$:

$$\max_{\mathbf{w} \in \mathbb{R}^d, \gamma \geq 0} \gamma \quad \text{subject to} \quad \text{margin}_i(\mathbf{w}) \geq \gamma \quad \text{for all } i = 1, \dots, n, \quad (1)$$

where

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

is the margin achieved by \mathbf{w} on the i th training example⁵ (\mathbf{x}_i, y_i) . The SVM gets its name from the fact that the solution (\mathbf{w}^*, γ^*) depends only on the set of training examples that achieve the minimum margin value, γ^* . These examples are known as the ‘support vectors’, and it is well-known that the weight vector \mathbf{w}^* can be written as a (non-negative) linear combination of the $y_i \mathbf{x}_i$ corresponding to support vectors. More precisely, the dual form of the solution expresses the weight vector $\mathbf{w}^* = \sum_{i=1}^n \alpha_i^* y_i \mathbf{x}_i$ in terms of dual variables $\alpha_1^*, \dots, \alpha_n^* \geq 0$. This constitutes a concise representation of the solution—just the list of non-zero dual variables α_i^* and corresponding data points. This remarkable property of the SVM is particularly important in its ‘kernelized’ extension [8, 39], 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 [19, 20, 43] 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 by a quantity that tends to zero as $n \rightarrow \infty$. Moreover, this sparsity in support vectors can be demonstrated in sufficiently low-dimensional settings using asymptotic arguments [10, 16, 30]. 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 [34] 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 [5, 7, 22, 32, 33, 35] to be applied to the SVM. Using this equivalence, Muthukumar *et al* [34] 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 [29, 47]. More importantly, the level of overparameterization considered there was only sufficiently, not necessarily, high enough for support vector proliferation.

⁵We 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 equation (2) in section 2.1.

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* [5]), 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.

1.1. 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 [19, 20, 43], and the modern high-dimensional regime [13, 34]. 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. [10, 16, 28, 30]). One particular result of note is that of Buhot and Gordon [10], who consider a linearly⁶ separable setting where the training data inputs are drawn i.i.d. 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 $n, d \rightarrow \infty$):

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

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 $\delta \rightarrow \infty$. 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 $\delta \rightarrow 0$. In these cases, the generalization error bounds based on support vectors no longer provide non-trivial guarantees.

Muthukumar *et al* [34] 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 [5]. 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.

⁶We 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.

Chatterji and Long [13] 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 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* [14] and the ν -SVM [40]. 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 does 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 [4, 41]. Indeed, this fact has motivated algorithmic developments for sparsifying the SVM solution (e.g. [11, 17, 23]). Second, under some general conditions on the data distribution, it is also shown for the ν -SVM [40, proposition 5] that as $n \rightarrow \infty$ 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 [2, 3, 21, 31, 49]. 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* [34, 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 [13, 34].

Our work bears some resemblance to the early work of Cover [15] 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 $\mathbf{w} \in \mathbb{R}^d$ such that $\text{margin}_i(\mathbf{w}) > 0$ for all $i = 1, \dots, n$,

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.

There have been several developments on support vector proliferation since the initial publication of our work. First, Ardeshir *et al* [1] strengthened our converse result under the independent features model. For isotropic features, they show that $d = \Omega(n \log n)$ is necessary for a constant probability of support vector proliferation. They also provide a converse result for anisotropic features in terms of effective dimensions. In the special case of standard Gaussian features, they find that the transition occurs around $d \sim 2n \log n$, and they bound the width of the transition. Ardeshir *et al* also give empirical evidence for the universality of support vector proliferation under broader classes of feature distributions. Independently, Wang and Thrampoulidis [45] showed that support vector proliferation also occurs under sufficient effective overparameterization under the Gaussian mixture model; Cao *et al* [12] further sharpened and generalized these results to general sub-Gaussian mixture models. Finally, Wang *et al* [46] considered the multiclass case and showed a high-probability equivalence between not only the one-vs-all SVM [37] and interpolation (which follows as a direct consequence of this work), but also the multiclass SVM [48]. While the dual of the multiclass SVM required a different and novel treatment, subsequent steps in their proof leverage, in part, the arguments that are provided in this paper.

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 $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n) \in \mathbb{R}^d \times \{-1, +1\}$. A homogeneous linear classifier is specified by a weight vector $\mathbf{w} \in \mathbb{R}^d$, so that the prediction of this classifier on $\mathbf{x} \in \mathbb{R}^d$ is given by the sign of $\mathbf{x}^\top \mathbf{w}$. The ambiguity of the sign when $\mathbf{x}^\top \mathbf{w} = 0$ is not important in our analysis.

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

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

The well-known Lagrangian dual of equation (2) can be written entirely in terms of the vector of labels $\mathbf{y} := (y_1, \dots, y_n) \in \mathbb{R}^n$ and the $n \times n$ Gram (or kernel) matrix \mathbf{K}

corresponding to $\mathbf{x}_1, \dots, \mathbf{x}_n$, i.e. $K_{i,j} := \mathbf{x}_i^\top \mathbf{x}_j$ for all $1 \leq i, j \leq n$:

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

Above, we use $\text{diag}(\cdot)$ to denote the diagonal matrix with diagonal entries taken from the vector-valued argument. An optimal solution $\boldsymbol{\alpha}^*$ to the dual problem in equation (3) corresponds to an optimal primal variable \mathbf{w}^* for the problem in equation (2) via the relation $\mathbf{w}^* = \sum_{i=1}^n \alpha_i^* y_i \mathbf{x}_i$. The support vectors are precisely the examples (\mathbf{x}_i, y_i) for which the corresponding α_i^* 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 $i = 1, \dots, n$. In terms of $\boldsymbol{\beta}$, the SVM dual problem from equation (3) becomes

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

An optimal solution $\boldsymbol{\beta}^*$ to this problem corresponds to an optimal primal variable \mathbf{w}^* via the relation $\mathbf{w}^* = \sum_{i=1}^n \beta_i^* \mathbf{x}_i$, and the support vectors are precisely the examples (\mathbf{x}_i, y_i) for which β_i^* 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 $\mathbf{K} \boldsymbol{\beta} = \mathbf{y}$. We refer to the version of the optimization problem in equation (4) without the n constraints as the *ridgeless regression problem*. Solutions to this problem have been extensively studied in recent years (e.g. [5, 7, 22, 27, 29, 35]). If a vector $\boldsymbol{\beta} \in \mathbb{R}^n$ satisfies both $\mathbf{K} \boldsymbol{\beta} = \mathbf{y}$ as well as the n constraints $y_i \beta_i \geq 0$ for all $i = 1, \dots, n$, then $\boldsymbol{\beta}$ is necessarily an optimal solution to the SVM dual problem from equation (4).

2.2. Data model

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

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

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

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

- (a) *Independent features:* \mathbf{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 $\mathbb{E}(e^{tz_{i,j}}) \leq e^{vt^2/2}$ for all $t \in \mathbb{R}$).
- (b) *Haar features:* \mathbf{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 $\mathbf{x}_1, \dots, \mathbf{x}_n$, the y_1, \dots, y_n are independent $\{-1, +1\}$ -valued random variables such that the conditional distribution of y_i depends only on \mathbf{x}_i for each $i = 1, \dots, n$. Formally:

$$y_i \perp (\mathbf{x}_1, y_1, \dots, \mathbf{x}_{i-1}, y_{i-1}, \mathbf{x}_{i+1}, y_{i+1}, \dots, \mathbf{x}_n, y_n) | \mathbf{x}_i.$$

Remarks. All of our results will assume $d \geq n$. The non-singularity of the kernel matrix $\mathbf{K} = \mathbf{Z} \text{diag}(\boldsymbol{\lambda}) \mathbf{Z}^\top$ will be important for our analysis. In the case of Haar features, setting $d \geq n$ ensures that the matrix \mathbf{Z} always has rank n , and hence the kernel matrix $\mathbf{K} = \mathbf{Z} \text{diag}(\boldsymbol{\lambda}) \mathbf{Z}^\top$ is always non-singular. In the case of independent features, if the distributions of the $z_{i,j}$ are continuous, then \mathbf{Z} has rank n almost surely, and hence again \mathbf{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 \mathbf{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.

- (a) *Generalized linear models (GLMs)*: $\Pr(y_i = 1 | \mathbf{x}_i) = g(\mathbf{x}_i^\top \mathbf{w})$ for some $\mathbf{w} \in \mathbb{R}^d$ and some function $g: \mathbb{R} \rightarrow [0, 1]$. 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* [9], where $g(t) = \mathbb{1}_{\{t > 0\}}$.
- (b) *Multi-index models*: $\Pr(y_i = 1 | \mathbf{x}_i) = h(\mathbf{W} \mathbf{x}_i)$ for some $k \in \mathbb{N}$, $\mathbf{W} \in \mathbb{R}^{k \times d}$, and $h: \mathbb{R}^k \rightarrow [0, 1]$. The case $k = 1$ corresponds to GLMs. Examples with $k \geq 2$ include the intersections of half-spaces models and certain neural networks [6, 24, 25].
- (c) *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 [15].

Our results in theorems 1 and 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 i.i.d. 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 $\mathbb{R}_{++} := \{x \in \mathbb{R} : x > 0\}$ denote the positive real numbers. For a vector $\mathbf{v} \in \mathbb{R}^n$, we let $\mathbf{v}_{\setminus i} \in \mathbb{R}^{n-1}$ denote the vector obtained from \mathbf{v} by omitting the i th coordinate. For a matrix $\mathbf{M} \in \mathbb{R}^{n \times d}$, we let $\mathbf{M}_{\setminus i} \in \mathbb{R}^{(n-1) \times d}$ denote the matrix obtained from \mathbf{M} by omitting the i th row. Sometimes, for a square matrix $\mathbf{M} \in \mathbb{R}^{n \times n}$, we will also use $\mathbf{M}_{\setminus i} \in \mathbb{R}^{(n-1) \times (n-1)}$ to denote the matrix obtained from \mathbf{M} by removing the i th row and column. We let \mathbf{e}_i denote the i th coordinate vector in \mathbb{R}^n . For a vector $\mathbf{v} \in \mathbb{R}^d$, we denote its p -norm by $\|\mathbf{v}\|_p = (\sum_{i=1}^d |v_i|^p)^{1/p}$. For a matrix $\mathbf{M} \in \mathbb{R}^{d \times d}$, we denote its $2 \rightarrow 2$ operator norm (i.e. largest singular value) by

$\|\mathbf{M}\|_{\text{op}} = \sup_{v \in \mathbb{R}^d: \|v\|_2 \leq 1} \|\mathbf{M}v\|_2$. Let $S^{d-1} := \{\mathbf{x} \in \mathbb{R}^d: \|\mathbf{x}\|_2 = 1\}$ denote the unit sphere in \mathbb{R}^d . If \mathbf{M} is a symmetric matrix, $\lambda_{\min}(\mathbf{M})$ denotes the smallest eigenvalue of \mathbf{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 $n \rightarrow \infty$. 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 section 4.

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

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

Observe that $d \geq d_2 \geq d_\infty$, 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 $r_0(\text{diag}(\boldsymbol{\lambda}))$ and $R_0(\text{diag}(\boldsymbol{\lambda}))$ studied by Bartlett *et al* [5]. They arise naturally from the tail behavior of certain linear combinations of χ^2 -random variables (see, e.g. [26]).

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 $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$ 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

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

We can compare this condition to that from the prior work of Muthukumar *et al* [34] in our setting with independent Gaussian features ($v = 1$). In the anisotropic setting

(i.e. general λ), 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 case⁷.

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 $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$ follow the model from section 2.2 with Haar features and model parameter $\lambda \in \mathbb{R}_{++}^d$, then the probability that every training example is a support vector is at least*

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

3.2. Weak converse

Our final main result gives a weak converse to theorem 1 in the case where the features are i.i.d. standard Gaussian and the labels are fixed.

Theorem 3. *Let the training data $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$ follow the model from section 2.2 with $\lambda = (1, \dots, 1)$, $\mathbf{z}_1, \dots, \mathbf{z}_n$ being i.i.d. standard Gaussian random vectors in \mathbb{R}^d , and $y_1, \dots, y_n \in \{\pm 1\}$ 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 A, 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 λ . These restrictions are related to the effective ranks studied by Bartlett *et al* [5]. The proof is similar to that of theorem 3, but also relies on a technical result from [5].

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

⁷ We remark that the result of Muthukumar *et al* [34] for the anisotropic case, in fact, holds for all (fixed) label vectors $\mathbf{y} \in \{-1, +1\}^n$ 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.

condition from theorem 3. As mentioned earlier, subsequent work of Ardeshir *et al* [1] closed this gap and showed that even for constant success probability, $d = \Omega(n \log n)$ is necessary.

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

- (a) Every training example is a support vector.
- (b) The vector $\boldsymbol{\beta} := \mathbf{K}^{-1} \mathbf{y}$ satisfies $y_i \beta_i > 0$ for all $i = 1, \dots, n$.
- (c) $y_i \mathbf{y}_i^\top (\mathbf{Z}_{\setminus i} \text{diag}(\boldsymbol{\lambda}) \mathbf{Z}_{\setminus i}^\top)^{-1} \mathbf{Z}_{\setminus i} \text{diag}(\boldsymbol{\lambda}) \mathbf{z}_i < 1$ for all $i = 1, \dots, n$.

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* [34], 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 theorems 1 and 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 [7, 35]) 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. $n/d \rightarrow \delta$ as $n, d \rightarrow \infty$). 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

On the proliferation of support vectors in high dimensions*

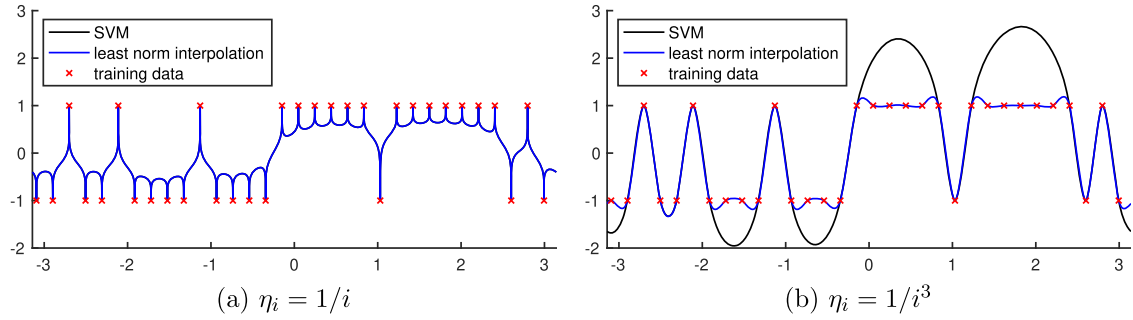


Figure 1. Plots of linear functions on top of trigonometric features of a scalar input variable that parameterizes the horizontal axis. (These plots originally appeared in [34].) The two linear functions are those given by the solution to the SVM optimization problem and the ridgeless regression problem (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 $t \mapsto (1, \sqrt{\eta_1} \cos(1 \cdot t), \sqrt{\eta_1} \sin(1 \cdot t), \dots, \sqrt{\eta_k} \cos(k \cdot t), \sqrt{\eta_k} \sin(k \cdot t)) \in \mathbb{R}^{2k+1}$ 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.

the form⁸

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

Here, \mathbf{w}^* is the solution to the SVM primal problem in equation (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

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

Since $\mathbb{E}[\text{tr}(\mathbf{K})]/n = \|\boldsymbol{\lambda}\|_1$, 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. [18]) 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 $1/\|\mathbf{w}^*\|_2$.

⁸Some 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.

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* [34] 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 [34, section 3.4] for further details. It suffices for our purposes to note that the main result of Muthukumar *et al* [34, 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 $n \rightarrow \infty$:

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

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

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

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* [34] is clear. In particular [34, 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

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

which has a non-trivial gap compared to equation (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. Proofs

This section gives the proofs of the main results, as well as the proof of the main technical lemma.

Throughout, we use the shorthand notations $\mathbf{\Lambda} := \text{diag}(\boldsymbol{\lambda})$ and $\mathbf{K}_{\setminus i} := \mathbf{Z}_{\setminus i} \mathbf{\Lambda} \mathbf{Z}_{\setminus i}^\top$ for each $i = 1, \dots, n$. Note that $\mathbf{K}_{\setminus i}$ is the same as $\mathbf{K} = \mathbf{Z} \text{diag}(\boldsymbol{\lambda}) \mathbf{Z}^\top$ except omitting both the i th row and the i th column (whereas $\mathbf{Z}_{\setminus i}$ only omits the i th row of \mathbf{Z}).

4.1. Proof of lemma 1

Recall that we assume \mathbf{K} and \mathbf{K}_{ni} 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 $\boldsymbol{\beta} = \mathbf{K}^{-1}\mathbf{y}$ satisfies

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

- (\Leftarrow) Assume $y_i\beta_i > 0$ for all $i \in [n]$. Recall that $\boldsymbol{\beta} = \mathbf{K}^{-1}\mathbf{y}$ is the unique optimal solution to the ridgeless regression problem (i.e. the problem in equation (4) without the n constraints). Since equation (6) holds, then $\boldsymbol{\beta}$ is dual-feasible as well, and so it is the unique optimal solution to the dual program, i.e. $\boldsymbol{\beta}^* = \boldsymbol{\beta}$. Moreover, $y_i\beta_i^* > 0 \implies \beta_i^* \neq 0$ for all $i \in [n]$, and so every training example is a support vector.
- (\Rightarrow) Assume every training example is a support vector, i.e. $\beta_i^* \neq 0$ for all $i \in [n]$ (so, in particular, $y_i\beta_i^* > 0$ for all $i \in [n]$). We shall write the solution \mathbf{w}^* to the primal problem from equation (2) as a linear combination of $\mathbf{x}_1, \dots, \mathbf{x}_n$ in two ways. The first way is in terms of the dual solution $\boldsymbol{\beta}^*$, i.e. $\mathbf{w}^* = \sum_{i=1}^n \beta_i^* \mathbf{x}_i$, which follows by strong duality. The second way comes via complementary slackness, which implies that \mathbf{w}^* satisfies every constraint in equation (2) with equality. In other words, \mathbf{w}^* solves

$$\begin{aligned} \min_{\mathbf{w} \in \mathbb{R}^d} \quad & \frac{1}{2} \|\mathbf{w}\|_2^2 \\ \text{subj. to } & \mathbf{x}_i^\top \mathbf{w} = y_i \quad \text{for all } i = 1, \dots, n. \end{aligned}$$

Since \mathbf{K} is non-singular by assumption, the solution is unique and is given by $\mathbf{X}^\top \mathbf{K}^{-1} \mathbf{y} = \mathbf{X}^\top \boldsymbol{\beta} = \sum_{i=1}^n \beta_i \mathbf{x}_i$, where $\mathbf{X} = [\mathbf{x}_1 | \dots | \mathbf{x}_n]^\top$. So we have $\mathbf{w}^* = \sum_{i=1}^n \beta_i^* \mathbf{x}_i = \sum_{i=1}^n \beta_i \mathbf{x}_i$. The non-singularity of \mathbf{K} also implies that $\mathbf{x}_1, \dots, \mathbf{x}_n$ are linearly independent, so we must have $\beta_i = \beta_i^* \neq 0$ for all $i \in [n]$, and thus equation (6) holds.

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

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

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

Observe that $y_1\beta_1 = y_1 \mathbf{e}_1^\top \mathbf{K}^{-1} \mathbf{y} = \mathbf{e}_1^\top \mathbf{K}^{-1} (y_1 \mathbf{y})$ is the inner product between the first row of \mathbf{K}^{-1} and $y_1 \mathbf{y}$. Therefore, by Cramer's rule, we have

$$y_1\beta_1 = y_1 \mathbf{e}_1^\top \mathbf{K}^{-1} \mathbf{y} = \frac{\det(\tilde{\mathbf{K}})}{\det(\mathbf{K})}$$

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

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

$$\tilde{\mathbf{K}} = \begin{bmatrix} 1 & y_1 \mathbf{y}_{\setminus 1}^\top \\ \mathbf{a} & \mathbf{K}_{\setminus 1} \end{bmatrix},$$

where $\mathbf{a} := \mathbf{Z}_{n1} \mathbf{\Lambda} \mathbf{z}_1$ and recall that \mathbf{K}_{n1} denotes the $(n-1) \times (n-1)$ matrix obtained by removing the first row and column from \mathbf{K} . Note that \mathbf{K}_{n1} is invertible by assumption and hence positive definite. Also, define

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

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

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

where $\mathbf{b} := y_1 \mathbf{y}_{n1}$. Therefore, $\det(\tilde{\mathbf{K}}) > 0$ iff $\det(\mathbf{K}_{\setminus 1} - \mathbf{a} \mathbf{b}^\top) > 0$.

By the matrix determinant lemma,

$$\det(\mathbf{K}_{\setminus 1} - \mathbf{a} \mathbf{b}^\top) = \det(\mathbf{K}_{\setminus 1}) (1 - \mathbf{b}^\top \mathbf{K}_{\setminus 1}^{-1} \mathbf{a}).$$

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

Connecting all of the equivalences and plugging-in for \mathbf{a} , \mathbf{b} , and \mathbf{K}_{n1} , we have shown that

$$y_1 \beta_1 > 0 \iff y_1 \mathbf{y}_{\setminus 1}^\top (\mathbf{Z}_{\setminus 1} \mathbf{\Lambda} \mathbf{Z}_{\setminus 1}^\top)^{-1} \mathbf{Z}_{\setminus 1} \mathbf{\Lambda} \mathbf{z}_1 < 1,$$

as required. This completes the proof of the lemma. \square

4.2. Proof of theorem 1

We fix $t = t(n, \mathbf{\Lambda}) > 0$ to a positive value depending on $\mathbf{\Lambda}$ and n that will be determined later. We define the following events:

- (a) For $i \in [n]$, \mathcal{B}_i is the event that \mathbf{K}_{ni} is non-singular and

$$y_i \mathbf{y}_{\setminus i}^\top \mathbf{K}_{\setminus i}^{-1} \mathbf{Z}_{\setminus i} \mathbf{\Lambda} \mathbf{z}_i \geq 1.$$

- (b) For $i \in [n]$, \mathcal{S}_i is the event that \mathbf{K}_{ni} is singular.

- (c) \mathcal{S} is the event that \mathbf{K} is singular.

- (d) $\mathcal{B} := \mathcal{S} \cup \bigcup_{i=1}^n (\mathcal{B}_i \cup \mathcal{S}_i)$.

Additionally, we define the event $\mathcal{E}_i(t)$, for every $i \in [n]$ and a given $t > 0$, that \mathbf{K}_{ni} is non-singular and

$$\left\| \mathbf{\Lambda} \mathbf{Z}_{\setminus i}^{\top} \mathbf{K}_{\setminus i}^{-1} \mathbf{y}_{\setminus i} \right\|_2^2 \geq \frac{1}{t}.$$

Note that if the event \mathcal{B} does not occur, then $\mathbf{Z} \mathbf{\Lambda} \mathbf{Z}^{\top}$ is non-singular, each \mathbf{K}_{ni} is non-singular, and

$$\mathbf{y}_i \mathbf{y}_i^{\top} \mathbf{K}_{\setminus i}^{-1} \mathbf{Z}_{\setminus i} \mathbf{\Lambda} \mathbf{z}_i < 1, \quad \text{for all } i = 1, \dots, n.$$

Hence, by lemma 1, if \mathcal{B} does not occur, then every training example is a support vector.

So, it suffices to upper-bound the probability of the event \mathcal{B} . We bound $\Pr(\mathcal{B})$ as follows:

$$\begin{aligned} \Pr(\mathcal{B}) &\leq \Pr(\mathcal{S}) + \sum_{i=1}^n \Pr(\mathcal{B}_i \cup \mathcal{S}_i) \\ &= \Pr(\mathcal{S}) + \sum_{i=1}^n (\Pr((\mathcal{B}_i \cap \mathcal{S}_i^c \cap \mathcal{E}_i(t)^c) \cup (\mathcal{S}_i \cap \mathcal{E}_i(t)^c)) + \Pr((\mathcal{B}_i \cup \mathcal{S}_i) \cap \mathcal{E}_i(t))) \\ &\leq \Pr(\mathcal{S}) + \sum_{i=1}^n (\Pr(\mathcal{B}_i | \mathcal{S}_i^c \cap \mathcal{E}_i(t)^c) \Pr(\mathcal{S}_i^c \cap \mathcal{E}_i(t)^c) \\ &\quad + \Pr(\mathcal{S}_i \cap \mathcal{E}_i(t)^c) + \Pr((\mathcal{B}_i \cup \mathcal{S}_i) \cap \mathcal{E}_i(t))) \\ &\leq \Pr(\mathcal{S}) + \sum_{i=1}^n (\Pr(\mathcal{B}_i | \mathcal{S}_i^c \cap \mathcal{E}_i(t)^c) + \Pr(\mathcal{S}_i) + \Pr(\mathcal{E}_i(t))). \end{aligned} \quad (7)$$

Above, the first two inequalities follow from the union bound, and the rest uses the law of total probability.

We first upper bound the probability of the singularity events in the following lemma.

Lemma 2. *We have*

$$\max\{\Pr(\mathcal{S}), \Pr(\mathcal{S}_1), \dots, \Pr(\mathcal{S}_n)\} \leq 2 \times 9^n \cdot \exp\left(-c \cdot \min\left\{\frac{d_2}{v^2}, \frac{d_{\infty}}{v}\right\}\right)$$

where $c > 0$ is the universal constant in the statement of lemma 8.

Proof. It suffices to bound $\Pr(\mathcal{S})$, since each \mathbf{K}_{ni} is a principal submatrix of \mathbf{K} , and hence $\lambda_{\min}(\mathbf{K}_{ni}) \geq \lambda_{\min}(\mathbf{K})$ for all $i \in [n]$. Observe that

$$\mathbf{Z} \mathbf{\Lambda} \mathbf{Z}^{\top} = \sum_{j=1}^d \lambda_j \mathbf{v}_j \mathbf{v}_j^{\top}$$

where \mathbf{v}_j is the j th column of \mathbf{Z} . Recall that the columns of \mathbf{Z} are independent, and so these vectors satisfy the conditions of lemma 8. Moreover, since $\mathbf{Z} \mathbf{\Lambda} \mathbf{Z}^{\top}$ is positive semi-definite, its singularity would require

$$\|\mathbf{Z}\mathbf{\Lambda}\mathbf{Z}^\top - \|\mathbf{\Lambda}\|_1 \mathbf{I}\|_2 \geq \|\mathbf{\Lambda}\|_1.$$

The probability of this latter event can be bounded by lemma 8 with $\tau = \|\mathbf{\Lambda}\|_1$, thereby giving the claimed bound on $\Pr(\mathcal{S})$. This completes the proof of the lemma. \square

The next lemma upper bounds the probability of the event \mathcal{B}_i conditioned on the non-singularity event \mathcal{S}_i and the complement of the event $\mathcal{E}_i(t)$.

Lemma 3. For any $t > 0$,

$$\Pr(\mathcal{B}_i | \mathcal{S}_i^c \cap \mathcal{E}_i(t)^c) \leq 2 \exp\left(-\frac{t}{2v}\right).$$

Proof. Let \mathcal{B}'_i be the event that $\mathbf{K}_{\setminus i}$ is non-singular and

$$\left| \mathbf{y}_{\setminus i}^\top \mathbf{K}_{\setminus i}^{-1} \mathbf{Z}_{\setminus i} \mathbf{\Lambda} \mathbf{z}_i \right| = \max \left\{ -\mathbf{y}_{\setminus i}^\top \mathbf{K}_{\setminus i}^{-1} \mathbf{Z}_{\setminus i} \mathbf{\Lambda} \mathbf{z}_i, \mathbf{y}_{\setminus i}^\top \mathbf{K}_{\setminus i}^{-1} \mathbf{Z}_{\setminus i} \mathbf{\Lambda} \mathbf{z}_i \right\} \geq 1.$$

Since $|y_i| = 1$, it follows that $\mathcal{B}_i \subseteq \mathcal{B}'_i$, so

$$\Pr(\mathcal{B}_i | \mathcal{S}_i^c \cap \mathcal{E}_i(t)^c) \leq \Pr(\mathcal{B}'_i | \mathcal{S}_i^c \cap \mathcal{E}_i(t)^c).$$

Conditional on the event $\mathcal{S}_i^c \cap \mathcal{E}_i(t)^c$, we have that $\mathbf{K}_{\setminus i}$ is non-singular and $\|\mathbf{\Lambda} \mathbf{Z}_{\setminus i}^\top \mathbf{K}_{\setminus i}^{-1} \mathbf{y}_{\setminus i}\|_2^2 \leq 1/t$. Since \mathbf{z}_i is independent of $\{(\mathbf{z}_j, y_j) : j \neq i\}$, it follows that

$$\mathbf{y}_{\setminus i}^\top \mathbf{K}_{\setminus i}^{-1} \mathbf{Z}_{\setminus i} \mathbf{\Lambda} \mathbf{z}_i = (\mathbf{\Lambda} \mathbf{Z}_{\setminus i}^\top \mathbf{K}_{\setminus i}^{-1} \mathbf{y}_{\setminus i})^\top \mathbf{z}_i$$

is (conditionally) sub-Gaussian with parameter at most $v \cdot \|\mathbf{\Lambda} \mathbf{Z}_{\setminus i}^\top \mathbf{K}_{\setminus i}^{-1} \mathbf{y}_{\setminus i}\|_2^2 \leq v/t$. Then, the standard sub-Gaussian tail bound gives us

$$\Pr(\mathcal{B}_i | \mathcal{S}_i^c \cap \mathcal{E}_i(t)^c) \leq \Pr(\mathcal{B}'_i | \mathcal{S}_i^c \cap \mathcal{E}_i(t)^c) \leq 2 \exp\left(-\frac{t}{2v}\right).$$

This completes the proof of the lemma. \square

Finally, the following lemma upper bounds the probability of the event $\mathcal{E}_i(t)$ for $t := d_\infty/2n$.

Lemma 4.

$$\Pr(\mathcal{E}_i(d_\infty/(2n))) \leq 2 \times 9^{n-1} \cdot \exp\left(-c \cdot \min\left\{\frac{d_2}{4v^2}, \frac{d_\infty}{v}\right\}\right)$$

where $c > 0$ is the universal constant from lemma 8.

Proof. Let $\mathcal{E}'_i(t)$ be the event that

$$\lambda_{\min}(\mathbf{K}_{\setminus i}) \leq n \|\mathbf{\Lambda}\|_\infty t.$$

Under \mathcal{S}_i^c , the matrix $\mathbf{K}_{\setminus i}$ is non-singular. We get

$$\begin{aligned}
\|\Lambda \mathbf{Z}_{\setminus i}^T \mathbf{K}_{\setminus i}^{-1} \mathbf{y}_{\setminus i}\|_2^2 &\leq \|\Lambda^{1/2}\|_{\text{op}}^2 \|\Lambda^{1/2} \mathbf{Z}_{\setminus i}^T \mathbf{K}_{\setminus i}^{-1} \mathbf{y}_{\setminus i}\|_2^2 \\
&= \|\boldsymbol{\lambda}\|_\infty \mathbf{y}_{\setminus i}^T \mathbf{K}_{\setminus i}^{-1} \mathbf{Z}_{\setminus i} \Lambda \mathbf{Z}_{\setminus i}^T \mathbf{K}_{\setminus i}^{-1} \mathbf{y}_{\setminus i} \\
&\leq n \|\boldsymbol{\lambda}\|_\infty \sup_{\mathbf{u} \in \mathbb{R}^{n-1} : \|\mathbf{u}\|_2=1} \mathbf{u}^T \mathbf{K}_{\setminus i}^{-1} \mathbf{u} \\
&= \frac{n \|\boldsymbol{\lambda}\|_\infty}{\lambda_{\min}(\mathbf{K}_{\setminus i})}.
\end{aligned}$$

It follows that $\mathcal{E}_i(t) \subseteq \mathcal{E}'_i(t)$. Observe that for $t := d_\infty/(2n)$, the event $\mathcal{E}'_i(t)$ is that where

$$\lambda_{\min}(\mathbf{K}_{\setminus i}) \leq \frac{1}{2} \|\boldsymbol{\lambda}\|_1.$$

Therefore (as in the proof of lemma 2), lemma 8 with $\tau = \|\boldsymbol{\lambda}\|_1/2$ implies that

$$\begin{aligned}
\Pr(\mathcal{E}'_i(d_\infty/(2n))) &= \Pr\left(\lambda_{\min}(\mathbf{K}_{\setminus i}) \leq \frac{1}{2} \|\boldsymbol{\lambda}\|_1\right) \\
&\leq 2 \times 9^{n-1} \cdot \exp\left(-c \cdot \min\left\{\frac{d_2}{4v^2}, \frac{d_\infty}{v}\right\}\right).
\end{aligned}$$

This completes the proof of the lemma. \square

Plugging the probability bounds from lemmas 2–4 (with $t = d_\infty/(2n)$) into equation (7) completes the proof of theorem 1. \square

4.3. Proof of theorem 2

The proof follows a similar sequence of steps to that of theorem 1 with slight differences in the events that we condition on. We first observe that $\frac{1}{\sqrt{d}} \mathbf{z}_i | (\mathbf{Z}_{\setminus i}, \mathbf{y}_{\setminus i})$ is a uniformly random unit vector in S^{d-1} restricted to the subspace orthogonal to the row space of \mathbf{Z}_{ni} . That is, it has the same (conditional) distribution as $\mathbf{B}_i \mathbf{u}_i$, where:

- (a) \mathbf{B}_i is a $d \times (n - d + 1)$ matrix whose columns form an orthonormal basis for the orthogonal complement of \mathbf{Z}_{ni} 's row space;
- (b) \mathbf{u}_i is a uniformly random unit vector in S^{d-n} .

As before, for every $i \in [n]$, we define the event \mathcal{B}_i that $\mathbf{K}_{\setminus i}$ is non-singular and

$$\mathbf{y}_i \mathbf{y}_{\setminus i}^T \mathbf{K}_{\setminus i}^{-1} \mathbf{Z}_{\setminus i} \Lambda \mathbf{z}_i \geq 1.$$

The Haar measure ensures that the matrices \mathbf{Z} and \mathbf{Z}_{ni} always have full row rank. Therefore, because $\Lambda \succ \mathbf{0}$, the matrices \mathbf{K} and \mathbf{K}_{ni} are always non-singular. So we do not need to worry about singularity (cf the events \mathcal{S} and \mathcal{S}_i). We accordingly consider the event $\mathcal{B} := \bigcup_{i=1}^n \mathcal{B}_i$. As before, we also define the event $\mathcal{E}_i(t)$ for every $i \in [n]$ and a given $t > 0$, that

$$\|\mathbf{B}_i^T \Lambda \mathbf{Z}_{\setminus i}^T \mathbf{K}_{\setminus i}^{-1} \mathbf{y}_{\setminus i}\|_2^2 \geq \frac{d - n + 1}{d} \cdot \frac{1}{t}.$$

By the union bound, we get

$$\begin{aligned}\Pr(\mathcal{B}) &\leq \sum_{i=1}^n \Pr(\mathcal{B}_i) \\ &\leq \sum_{i=1}^n \Pr(\mathcal{B}_i | \mathcal{E}_i(t)^c) + \Pr(\mathcal{E}_i(t)),\end{aligned}$$

and so we need to upper bound the probabilities $\Pr(\mathcal{B}_i | \mathcal{E}_i(t)^c)$ and $\Pr(\mathcal{E}_i(t))$ for every $i \in [n]$.

The following lemma upper bounds $\Pr(\mathcal{B}_i | \mathcal{E}_i(t)^c)$, and is analogous to lemma 3 in the proof of theorem 1.

Lemma 5. *For any $t > 0$, we have*

$$\Pr(\mathcal{B}_i | \mathcal{E}_i(t)^c) \leq 2 \exp(-t).$$

Proof. First, as discussed above, we have

$$\begin{aligned}\Pr\left(y_i \mathbf{y}_{\setminus i}^\top \mathbf{K}_{\setminus i}^{-1} \mathbf{Z}_{\setminus i} \boldsymbol{\Lambda} \mathbf{z}_i \geq 1\right) &= \Pr\left(\sqrt{d} \cdot y_i \mathbf{y}_{\setminus i}^\top \mathbf{K}_{\setminus i}^{-1} \mathbf{Z}_{\setminus i} \boldsymbol{\Lambda} \mathbf{B}_i \mathbf{u}_i \geq 1\right) \\ &\leq \Pr\left(\sqrt{d} \left| (\mathbf{B}_i^\top \boldsymbol{\Lambda} \mathbf{Z}_{\setminus i}^\top \mathbf{K}_{\setminus i}^{-1} \mathbf{y}_{\setminus i})^\top \mathbf{u}_i \right| \geq 1\right).\end{aligned}$$

Moreover, \mathbf{u}_i is independent of $\mathbf{Z}_{\setminus i}$, and as established in lemma 9, the random vector \mathbf{u}_i is sub-Gaussian with parameter at most $\mathcal{O}(1/(d-n+1))$. Therefore, $\sqrt{d} \cdot (\mathbf{B}_i^\top \boldsymbol{\Lambda} \mathbf{Z}_{\setminus i}^\top \mathbf{K}_{\setminus i}^{-1} \mathbf{y}_{\setminus i})^\top \mathbf{u}_i$ is conditionally sub-Gaussian with parameter at most $\frac{d}{d-n+1} \cdot \|\mathbf{B}_i^\top \boldsymbol{\Lambda} \mathbf{Z}_{\setminus i}^\top \mathbf{K}_{\setminus i}^{-1} \mathbf{y}_{\setminus i}\|_2^2 \leq \frac{1}{t}$. Here, the last inequality follows because we have conditioned on $\mathcal{E}_i(t)^c$. Therefore, the standard sub-Gaussian tail bound gives us

$$\Pr(\mathcal{B}_i | \mathcal{E}_i(t)^c) \leq 2 \exp(-t).$$

The next lemma upper bounds $\Pr(\mathcal{E}_i(t))$ for $t := \frac{d-n+1}{d} \cdot \frac{d_\infty}{2n}$, and is analogous to lemma 4 in the proof of theorem 1. \square

Lemma 6. *We have*

$$\Pr\left(\mathcal{E}_i\left(\frac{d-n+1}{d} \cdot \frac{d_\infty}{2n}\right)\right) \leq \exp(-c_1 \cdot d) + 2 \times 9^n \cdot \exp(-c_2 \cdot \min\{d_2, d_\infty\})$$

where $c_1 > 0$ and $c_2 > 0$ are universal constants.

Proof. We get

$$\begin{aligned}\|\mathbf{B}_i^\top \boldsymbol{\Lambda} \mathbf{Z}_{\setminus i}^\top \mathbf{K}_{\setminus i}^{-1} \mathbf{y}_{\setminus i}\|_2^2 &\leq \|\mathbf{B}_i^\top\|_2^2 \cdot \|\boldsymbol{\Lambda} \mathbf{Z}_{\setminus i}^\top \mathbf{K}_{\setminus i}^{-1} \mathbf{y}_{\setminus i}\|_2^2 \\ &= \|\boldsymbol{\Lambda} \mathbf{Z}_{\setminus i}^\top \mathbf{K}_{\setminus i}^{-1} \mathbf{y}_{\setminus i}\|_2^2 \\ &\leq \frac{n \|\boldsymbol{\Lambda}\|_\infty}{\lambda_{\min}(\mathbf{K}_{\setminus i})},\end{aligned}$$

where we used the fact that \mathbf{B}_i has orthonormal columns, and the last inequality follows by an identical argument to the proof of lemma 4. We will show in particular that

$$\Pr\left(\lambda_{\min}(\mathbf{K}_{\setminus i}) \geq \frac{1}{2}\|\boldsymbol{\lambda}\|_1\right) \geq 1 - \exp(-c_1 \cdot d) - 2 \times 9^{n-1} \cdot \exp(-c_2 \cdot \min\{d_2, d_\infty\}). \quad (8)$$

Given equation (8), we can complete the proof of lemma 6. This is because we get

$$\|\mathbf{B}_i^\top \boldsymbol{\Lambda} \mathbf{Z}_{\setminus i}^\top \mathbf{K}_{\setminus i}^{-1} \mathbf{y}_{\setminus i}\|_2^2 \leq \frac{2n\|\boldsymbol{\lambda}\|_\infty}{\|\boldsymbol{\lambda}\|_1} = \frac{2n}{d_\infty} = \frac{d-n+1}{d} \cdot \frac{1}{t}$$

for

$$t := \frac{d-n+1}{d} \cdot \frac{d_\infty}{2n}.$$

We complete the proof by proving equation (8). Let $\mathbf{S} \in \mathbb{R}^{m \times d}$ be a random matrix with i.i.d. standard Gaussian entries with $m := n-1$, and let the singular value decomposition of \mathbf{S} be $\mathbf{S} = \mathbf{V} \boldsymbol{\Lambda}_S \mathbf{U}^\top$ where $\mathbf{V} \in \mathbb{R}^{m \times m}$ and $\mathbf{U} \in \mathbb{R}^{d \times m}$ are orthonormal matrices. Then, it is well-known that $\sqrt{d} \cdot \mathbf{U}^\top$ follows the same distribution as $\mathbf{Z}_{\setminus i}$, and hence $\lambda_{\min}(\mathbf{K}_{\setminus i})$ has the same distribution as $d \cdot \lambda_{\min}(\mathbf{U}^\top \boldsymbol{\Lambda} \mathbf{U})$. Moreover,

$$\begin{aligned} d \cdot \lambda_{\min}(\mathbf{U}^\top \boldsymbol{\Lambda} \mathbf{U}) &= \min_{\mathbf{v} \in \mathbb{R}^n, \|\mathbf{v}\|_2=1} \mathbf{v}^\top \boldsymbol{\Lambda}_S^{-1} \mathbf{V}^\top \mathbf{V} \boldsymbol{\Lambda}_S \mathbf{U}^\top \boldsymbol{\Lambda} \mathbf{U} \boldsymbol{\Lambda}_S \mathbf{V}^\top \mathbf{V} \boldsymbol{\Lambda}_S^{-1} \mathbf{v} \\ &\geq \frac{d}{\|\boldsymbol{\Lambda}_S\|_{\text{op}}^2} \cdot \min_{\mathbf{v} \in \mathbb{R}^n, \|\mathbf{v}\|_2=1} \mathbf{v}^\top \mathbf{S} \boldsymbol{\Lambda} \mathbf{S}^\top \mathbf{v} \\ &= \frac{d}{\|\boldsymbol{\Lambda}_S\|_{\text{op}}^2} \cdot \lambda_{\min}(\mathbf{S} \boldsymbol{\Lambda} \mathbf{S}^\top). \end{aligned}$$

By classical operator norm tail bounds on Gaussian random matrices (e.g. [44], corollary 5.35), we note that $\|\boldsymbol{\Lambda}_S\|_2^2 \leq \frac{3}{2}d$ with probability at least $1 - \exp(-c_1 \cdot d)$. Now, we note that the matrix $\mathbf{S} \boldsymbol{\Lambda} \mathbf{S}^\top := \sum_{j=1}^d \lambda_j \mathbf{s}_j \mathbf{s}_j^\top$ where the \mathbf{s}_j 's are i.i.d. standard Gaussian random vectors in \mathbb{R}^n . So, we directly substitute lemma 8 with $\tau := \frac{1}{4}\|\boldsymbol{\lambda}\|_1$, and get $\lambda_{\min}(\mathbf{S} \boldsymbol{\Lambda} \mathbf{S}^\top) \geq \frac{3}{4}\|\boldsymbol{\lambda}\|_1$ with probability at least $1 - 2 \times 9^m \cdot \exp(-c_2 \cdot \min\{d_2, d_\infty\})$. Putting both of these inequalities together directly gives us equation (8) with the desired probability bound, and completes the proof. \square

Finally, putting the high probability statements of lemmas 5 and 6 together completes the proof of theorem 2.

4.4. Proof of theorem 3

By lemma 1, our task is equivalent to lower-bounding the probability that there exists $i \in [n]$ such that $y_i \mathbf{y}_{\setminus i}^\top (\mathbf{Z}_{\setminus i} \mathbf{Z}_{\setminus i}^\top)^{-1} \mathbf{Z}_{\setminus i} \mathbf{z}_i \geq 1$. This event is the union of n (possibly overlapping) events, and hence its probability is at least the probability of one of the events, say, the first one:

$$\Pr\left(\exists i \in [n] \text{ s.t. } y_i \mathbf{y}_{\setminus i}^\top \mathbf{K}_{\setminus i}^{-1} \mathbf{Z}_{\setminus i} \mathbf{z}_i \geq 1\right) \geq \Pr\left(y_1 \mathbf{y}_{\setminus 1}^\top \mathbf{K}_{\setminus 1}^{-1} \mathbf{Z}_{\setminus 1} \mathbf{z}_1 \geq 1\right).$$

Because \mathbf{z}_1 is a standard Gaussian random vector independent of $\mathbf{Z}_{\setminus 1}$, the conditional distribution of $y_1 \mathbf{y}_{\setminus 1}^\top \mathbf{K}_{\setminus 1}^{-1} \mathbf{Z}_{\setminus 1} \mathbf{z}_1 | \mathbf{Z}_{\setminus 1}$ is Gaussian with mean zero and variance $\sigma^2 := \|\mathbf{Z}_{\setminus 1}^\top \mathbf{K}_{\setminus 1}^{-1} \mathbf{y}_{\setminus 1}\|_2^2$. Therefore, for any $t > 0$, we have

$$\begin{aligned} \Pr(y_1 \mathbf{y}_{\setminus 1}^\top \mathbf{K}_{\setminus 1}^{-1} \mathbf{Z}_{\setminus 1} \mathbf{z}_1 \geq 1) &= \mathbb{E}[\Pr(\sigma g \geq 1 | \sigma)] \quad (\text{where } g \sim \mathcal{N}(0, 1), \text{ } \underline{g} \perp \underline{\mathbf{z}}) \\ &= \mathbb{E}[\Phi(-1/\sigma)] \\ &\geq \mathbb{E}[\Phi(-1/\sigma) | \sigma^2 \geq 1/t] \Pr(\sigma^2 \geq 1/t) \\ &\geq \Phi(-\sqrt{t}) \cdot \Pr(\mathcal{E}_1(t)), \end{aligned}$$

where Φ is the standard Gaussian cumulative distribution function, and $\mathcal{E}_1(t)$ is the event that

$$\sigma^2 = \mathbf{y}_{\setminus 1}^\top \mathbf{K}_{\setminus 1}^{-1} \mathbf{Z}_{\setminus 1} \mathbf{Z}_{\setminus 1}^\top \mathbf{K}_{\setminus 1}^{-1} \mathbf{y}_{\setminus 1} = \mathbf{y}_{\setminus 1}^\top \mathbf{K}_{\setminus 1}^{-1} \mathbf{y}_{\setminus 1} \geq \frac{1}{t}$$

(as in the proofs of theorems 1 and 2). We now lower-bound the probability of $\mathcal{E}_1(t)$. Observe that the $(n-1) \times (n-1)$ random matrix $\mathbf{K}_{\setminus 1} = \mathbf{Z}_{\setminus 1} \mathbf{Z}_{\setminus 1}^\top$ follows a Wishart distribution with identity scale matrix and d degrees-of-freedom. Moreover, by the rotational symmetry of the standard Gaussian distribution, the random variable $\mathbf{y}_{\setminus 1}^\top \mathbf{K}_{\setminus 1}^{-1} \mathbf{y}_{\setminus 1}$ has the same distribution as that of $(\sqrt{n-1} \mathbf{e}_1)^\top \mathbf{K}_{\setminus 1}^{-1} (\sqrt{n-1} \mathbf{e}_1) = (n-1) \mathbf{e}_1^\top \mathbf{K}_{\setminus 1}^{-1} \mathbf{e}_1$. It is known that $1/\mathbf{e}_1^\top \mathbf{K}_{\setminus 1}^{-1} \mathbf{e}_1$ follows a χ^2 distribution with $d - (n-2)$ degrees-of-freedom; we denote its cumulative distribution function by F_{d-n+2} . Therefore,

$$\Pr(\mathcal{E}_1(t)) = F_{d-n+2}(t(n-1)).$$

So, we have shown that

$$\Pr(y_1 \mathbf{y}_{\setminus 1}^\top \mathbf{K}_{\setminus 1}^{-1} \mathbf{Z}_{\setminus 1} \mathbf{z}_1 \geq 1) \geq \sup_{t \geq 0} \Phi(-\sqrt{t}) \cdot F_{d-n+2}(t(n-1)).$$

For $t := \frac{d-n+4+2\sqrt{d-n+2}}{n-1}$, we obtain $F_{d-n+2}(t) \geq 1 - 1/e$ by a standard χ^2 tail bound [26, lemma 1]. In this case, we obtain

$$\Pr(y_1 \mathbf{y}_{\setminus 1}^\top \mathbf{K}_{\setminus 1}^{-1} \mathbf{Z}_{\setminus 1} \mathbf{z}_1 \geq 1) \geq \Phi\left(-\sqrt{\frac{d-n+4+2\sqrt{d-n+2}}{n-1}}\right) \cdot \left(1 - \frac{1}{e}\right)$$

as claimed. \square

Acknowledgments

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. D H acknowledges partial support from NSF awards CCF-1740833 and IIS-1815697, a Sloan Research Fellowship, and a Google Faculty Award. V M acknowledges partial support from a Simons-Berkeley Research Fellowship, support of the ML4Wireless center member companies and NSF Grants AST-144078 and ECCS-1343398. J X was supported by a Cheung-Kong Graduate School of Business Fellowship as a PhD student at Columbia University during this project.

Appendix A. Anisotropic version of theorem 3

Below, we give a version of theorem 3 that applies to certain anisotropic settings, depending on some conditions on λ .

Theorem 4. *There are absolute constants $c > 0$ and $c' > 0$ such that the following hold. Let the training data $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$ follow the model from section 2.2, with $\mathbf{z}_1, \dots, \mathbf{z}_n$ being i.i.d. standard Gaussian random vectors in \mathbb{R}^d , and $y_1, \dots, y_n \in \{\pm 1\}$ being arbitrary but fixed (i.e. non-random) values. Assume $d > n$ and that there exists $k \in \mathbb{N}$ and $b > 1$ such that $k < (n-1)/c$ and*

$$\frac{\sum_{j=k+1}^d \lambda_j}{\lambda_{k+1}} \leq b(n-1)$$

where $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$. Then the probability that at least one training example is not a support vector is at least

$$c' \cdot \Phi \left(-\sqrt{\frac{2cb^2(n-1)}{k+1}} \right) \cdot (1 - 10e^{-(n-1)/c}),$$

where Φ is the standard Gaussian cumulative distribution function.

Note that the probability bound in theorem 4 is at least a positive constant for sufficiently large n provided that the (k, b) obtained as a function of λ satisfy $k+1 \geq c''b^2(n-1)$ for some absolute constant $c'' > 0$.

Proof. The proof begins in the same way as in that of theorem 3. Using the same arguments, we obtain the following lower bound:

$$\begin{aligned} \Pr(\exists i \in [n] \text{ s.t. } y_i \mathbf{y}_i^\top \mathbf{K}_{\setminus i}^{-1} \mathbf{Z}_{\setminus i} \Lambda \mathbf{z}_i \geq 1) &\geq \Pr(y_1 \mathbf{y}_1^\top \mathbf{K}_{\setminus 1}^{-1} \mathbf{Z}_{\setminus 1} \Lambda \mathbf{z}_1 \geq 1) \\ &\geq \Phi(-\sqrt{t}) \cdot \Pr(\mathcal{E}_1(t)) \end{aligned} \quad (9)$$

where $\mathcal{E}_1(t)$ is the event that

$$\left\| \Lambda \mathbf{Z}_{\setminus 1}^\top \mathbf{K}_{\setminus 1}^{-1} \mathbf{y}_{\setminus 1} \right\|_2^2 \geq \frac{1}{t}.$$

We next focus on lower-bounding the probability of $\mathcal{E}_1(t)$. (This part is more involved than in the proof of theorem 3.) Observe that the (rotationally invariant) distribution of $\mathbf{Z}_{\setminus 1}$ is the same as that of $\mathbf{Q}\mathbf{Z}_{\setminus 1}$, where \mathbf{Q} is a uniformly random $(n-1) \times (n-1)$

orthogonal matrix independent of $\mathbf{Z}_{\setminus 1}$. Therefore, $\Lambda \mathbf{Z}_{\setminus 1}^\top \mathbf{K}^{-1} \mathbf{y}_{\setminus 1}$ has the same distribution as

$$\begin{aligned} \Lambda (\mathbf{Q} \mathbf{Z}_{\setminus 1})^\top (\mathbf{Q} \mathbf{Z}_{\setminus 1} \Lambda \mathbf{Z}_{\setminus 1}^\top \mathbf{Q}^\top)^{-1} \mathbf{y}_{\setminus 1} &= \Lambda \mathbf{Z}_{\setminus 1}^\top \mathbf{Q}^\top \mathbf{Q} (\mathbf{Z}_{\setminus 1} \Lambda \mathbf{Z}_{\setminus 1}^\top)^{-1} \mathbf{Q}^\top \mathbf{y}_{\setminus 1} \\ &= \sqrt{n-1} \Lambda \mathbf{Z}_{\setminus 1}^\top \mathbf{K}_{\setminus 1}^{-1} \mathbf{u} \end{aligned}$$

where $\mathbf{u} := \mathbf{Q}^\top \mathbf{y}_{\setminus 1} / \sqrt{n-1}$ is a uniformly random unit vector, independent of $\mathbf{Z}_{\setminus 1}$. Letting $\mathbf{M} := \Lambda \mathbf{Z}_{\setminus 1}^\top \mathbf{K}_{\setminus 1}^{-1}$, we can thus lower-bound the probability of $\mathcal{E}_1(t)$ using

$$\begin{aligned} \Pr(\mathcal{E}_1(t)) &= \Pr(\|\sqrt{n-1} \mathbf{M} \mathbf{u}\|_2^2 > 1/t) \\ &\geq \Pr(\|\sqrt{n-1} \mathbf{M} \mathbf{u}\|_2^2 > 1/t \mid \text{tr}(\mathbf{M}^\top \mathbf{M}) \geq 2/t) \cdot \Pr(\text{tr}(\mathbf{M}^\top \mathbf{M}) \geq 2/t). \end{aligned} \quad (10)$$

We lower-bound each of the probabilities on the right-hand side of equation (10).

We begin with the first probability in equation (10), which we handle for arbitrary $t > 0$. By the Paley–Zygmund inequality, we have

$$\Pr\left(\|\sqrt{n-1} \mathbf{M} \mathbf{u}\|_2^2 > \frac{1}{2} \mathbb{E}[\|\sqrt{n-1} \mathbf{M} \mathbf{u}\|_2^2] \mid \mathbf{Z}_{\setminus 1}\right) \geq \frac{1}{4} \cdot \frac{\mathbb{E}[\|\sqrt{n-1} \mathbf{M} \mathbf{u}\|_2^2]^2}{\mathbb{E}[\|\sqrt{n-1} \mathbf{M} \mathbf{u}\|_2^4]}. \quad (11)$$

Since $\sqrt{n-1} \mathbf{u}$ is isotropic, we have

$$\mathbb{E}[\|\sqrt{n-1} \mathbf{M} \mathbf{u}\|_2^2 \mid \mathbf{Z}_{\setminus 1}] = (n-1) \text{tr}(\mathbf{M}^\top \mathbf{M} \mathbb{E}[\mathbf{u} \mathbf{u}^\top]) = \text{tr}(\mathbf{M}^\top \mathbf{M}).$$

Furthermore, by lemma 9,

$$\mathbb{E}[\|\sqrt{n-1} \mathbf{M} \mathbf{u}\|_2^4 \mid \mathbf{Z}_{\setminus 1}] \leq C \text{tr}(\mathbf{M}^\top \mathbf{M})^2$$

for some universal constant $C > 0$. Therefore, plugging back into equation (11), we obtain

$$\Pr\left(\|\sqrt{n-1} \mathbf{M} \mathbf{u}\|_2^2 > \frac{1}{2} \text{tr}(\mathbf{M}^\top \mathbf{M}) \mid \mathbf{Z}_{\setminus 1}\right) \geq \frac{1}{4} \cdot \frac{\text{tr}(\mathbf{M}^\top \mathbf{M})^2}{C \text{tr}(\mathbf{M}^\top \mathbf{M})^2} = \frac{1}{4C}.$$

Thus we also have the following for arbitrary $t > 0$:

$$\Pr(\|\sqrt{n-1} \mathbf{M} \mathbf{u}\|_2^2 > 1/t \mid \text{tr}(\mathbf{M}^\top \mathbf{M}) \geq 2/t) \geq \frac{1}{4C}. \quad (12)$$

We next consider the second probability in equation (10), namely $\Pr(\text{tr}(\mathbf{M}^\top \mathbf{M}) \geq 2/t)$. Recall that we assume there exists $k < (n-1)/c$ and $b > 1$ such that

$$\frac{\sum_{j=k+1}^d \lambda_j}{\lambda_{k+1}} \leq b(n-1). \quad (13)$$

We claim that for $t := \frac{2cb^2(n-1)}{k+1}$,

$$\Pr\left(\operatorname{tr}(\mathbf{M}^\top \mathbf{M}) \geq \frac{2}{t}\right) \geq 1 - 10e^{-(n-1)/c}. \quad (14)$$

Indeed, this claim follows from lemma 16 of [5], where their matrix \mathbf{C} is our matrix $\mathbf{M}^\top \mathbf{M}$, except our matrix is $(n-1) \times (n-1)$ instead of $n \times n$, and their matrix $\mathbf{\Sigma}$ is our matrix $\mathbf{\Lambda}$; see the definitions in their lemma 8. The universal constant $c > 0$ in their lemma is the same as ours, and equation (13) is precisely their condition $r_k(\mathbf{\Sigma}) < b(n-1)$ (with the same k and b). Therefore, the conclusion of their lemma implies, in our notation, that with probability at least $1 - 10e^{-(n-1)/c}$,

$$\operatorname{tr}(\mathbf{M}^\top \mathbf{M}) \geq \frac{k+1}{cb^2(n-1)} = \frac{2}{t}.$$

This proves the claimed probability bound.

We conclude from equations (9), (10), (12) and (14) that the probability that at least one training example is not a support vector is bounded below by

$$\Phi\left(-\sqrt{\frac{2cb^2(n-1)}{k+1}}\right) \cdot \frac{1}{4C} \cdot (1 - 10e^{-(n-1)/c})$$

as claimed. \square

Appendix B. Probabilistic inequalities

Lemma 7. *Let $\mathbf{M} \in \mathbb{R}^{n \times n}$ be a symmetric matrix, and let \mathcal{N} be an ϵ -net of S^{n-1} with respect to the Euclidean metric for some $\epsilon < 1/2$, then*

$$\|\mathbf{M}\|_2 \leq \frac{1}{1-2\epsilon} \max_{\mathbf{u} \in \mathcal{N}} |\mathbf{u}^\top \mathbf{M} \mathbf{u}|.$$

Proof. See [44, lemma 5.4]. \square

Lemma 8. *There is a universal constant $c > 0$ such that the following holds. Let $\lambda_1, \dots, \lambda_d > 0$ be given. Let $\mathbf{v}_1, \dots, \mathbf{v}_d$ be independent random vectors taking values in \mathbb{R}^n such that, for some $v > 0$,*

$$\mathbb{E}(\mathbf{v}_j) = \mathbf{0}, \mathbb{E}(\mathbf{v}_j \mathbf{v}_j^\top) = \mathbf{I}_n, \mathbb{E}(\exp(\mathbf{u}^\top \mathbf{v}_j)) \leq \exp(v \|\mathbf{u}\|_2^2 / 2) \quad \text{for all } \mathbf{u} \in \mathbb{R}^n$$

for all $j = 1, \dots, d$. For any $\tau > 0$,

$$\Pr\left(\left\|\sum_{j=1}^d \lambda_j \mathbf{v}_j \mathbf{v}_j^\top - \|\mathbf{\lambda}\|_1 \mathbf{I}_n\right\|_2 \geq \tau\right) \leq 2 \times 9^n \cdot \exp\left(-c \cdot \min\left\{\frac{\tau^2}{v^2 \|\mathbf{\lambda}\|_2^2}, \frac{\tau}{v \|\mathbf{\lambda}\|_\infty}\right\}\right).$$

where $\|\mathbf{\lambda}\|_1 := \sum_{j=1}^d \lambda_j$, $\|\mathbf{\lambda}\|_2^2 := \sum_{j=1}^d \lambda_j^2$, and $\|\mathbf{\lambda}\|_\infty := \max_{j \in [d]} \lambda_j$.

Proof. Let \mathcal{N} be an $(1/4)$ -net of S^{n-1} with respect to the Euclidean metric. A standard volume argument of Pisier [36] allows a choice of \mathcal{N} with $|\mathcal{N}| \leq 9^n$. By lemma 7, we

have for any $t > 0$,

$$\Pr\left(\left\|\sum_{j=1}^d \lambda_j \mathbf{v}_j \mathbf{v}_j^\top - \|\boldsymbol{\lambda}\|_1 \mathbf{I}_n\right\|_2 \geq \tau\right) \leq \Pr\left(\max_{\mathbf{u} \in \mathcal{N}} \left|\sum_{j=1}^d \lambda_j (\mathbf{u}^\top \mathbf{v}_j)^2 - \|\boldsymbol{\lambda}\|_1\right| \geq \tau/2\right).$$

Next, observe that for any $\mathbf{u} \in S^{n-1}$, the random variables $\mathbf{u}^\top \mathbf{v}_1, \dots, \mathbf{u}^\top \mathbf{v}_d$ are independent random variables, each with mean-zero, unit variance, and sub-Gaussian with parameter v . By the Hanson–Wright inequality of [38] and a union bound, there exists a universal constant $c > 0$ such that, for any unit vector $\mathbf{u} \in S^{n-1}$ and any $\tau > 0$,

$$\begin{aligned} & \Pr\left(\max_{\mathbf{u} \in \mathcal{N}} \left|\sum_{j=1}^d \lambda_j (\mathbf{u}^\top \mathbf{v}_j)^2 - \|\boldsymbol{\lambda}\|_1\right| \geq \tau/2\right) \\ & \leq 2 \times 9^n \cdot \exp\left(-c \cdot \min\left\{\frac{\tau^2}{v^2 \|\boldsymbol{\lambda}\|_2^2}, \frac{\tau}{v \|\boldsymbol{\lambda}\|_\infty}\right\}\right). \end{aligned}$$

The claim follows. \square

Lemma 9. Let $\boldsymbol{\theta}$ be a uniformly random unit vector in S^{m-1} . For any unit vector $\mathbf{u} \in S^{m-1}$, the random variable $\mathbf{u}^\top \boldsymbol{\theta}$ is sub-Gaussian with parameter $v = O(1/m)$. Moreover, for any matrix $\mathbf{M} \in \mathbb{R}^{m \times m}$, we have

$$\mathbb{E}[\|\mathbf{M}\boldsymbol{\theta}\|_2^4] \leq \frac{C}{m^2} \text{tr}(\mathbf{M}^\top \mathbf{M})^2$$

where $C > 0$ is a universal constant.

Proof. Let L be a χ random variable with m degrees-of-freedom, independent of $\boldsymbol{\theta}$, so the distribution of $\mathbf{z} := L\boldsymbol{\theta}$ is the standard Gaussian in \mathbb{R}^m . Let $\mu := \mathbb{E}[L] = \mathbb{E}[L|\boldsymbol{\theta}] = \sqrt{2} \frac{\Gamma((m+1)/2)}{\Gamma(m/2)} = \Omega(\sqrt{m})$. By Jensen's inequality, for any $t \in \mathbb{R}$,

$$\begin{aligned} \mathbb{E}[\exp(t \mathbf{u}^\top \boldsymbol{\theta})] &= \mathbb{E}\left[\exp\left(\left(\frac{t}{\mu} \mathbf{u}\right)^\top (\mathbb{E}[L|\boldsymbol{\theta}]\boldsymbol{\theta})\right)\right] \\ &\leq \mathbb{E}\left[\exp\left(\left(\frac{t}{\mu} \mathbf{u}\right)^\top (L\boldsymbol{\theta})\right)\right] \\ &= \mathbb{E}\left[\exp\left(\left(\frac{t}{\mu} \mathbf{u}\right)^\top \mathbf{z}\right)\right] \\ &= \exp\left(\frac{t^2}{2\mu^2}\right). \end{aligned}$$

It follows that $\mathbf{u}^\top \boldsymbol{\theta}$ is sub-Gaussian with parameter $v = 1/\mu^2 = O(1/m)$.

Similarly, again by Jensen's inequality,

$$\begin{aligned}\mu^4 \cdot \mathbb{E}[\|\mathbf{M}\boldsymbol{\theta}\|_2^4] &= \mathbb{E}[\mathbb{E}[L|\boldsymbol{\theta}]^4 \|\mathbf{M}\boldsymbol{\theta}\|_2^4] \\ &\leq \mathbb{E}[L^4 \|\mathbf{M}\boldsymbol{\theta}\|_2^4] \\ &= \mathbb{E}[\|\mathbf{M}\mathbf{z}\|_2^4].\end{aligned}$$

Furthermore, a direct computation shows that

$$\begin{aligned}\mathbb{E}[\|\mathbf{M}\mathbf{z}\|_2^4] &= 2 \operatorname{tr}((\mathbf{M}^\top \mathbf{M})^2) + \operatorname{tr}(\mathbf{M}^\top \mathbf{M})^2 \\ &\leq 3 \operatorname{tr}(\mathbf{M}^\top \mathbf{M})^2.\end{aligned}$$

The conclusion follows since $\mu^4 = \Omega(m^2)$. □

References

- [1] Ardeshtir N, Sanford C and Hsu D 2021 Support vector machines and linear regression coincide with very high-dimensional features *Advances in Neural Information Processing Systems* vol 34 pp 4907–18
- [2] Bartlett P and Shawe-Taylor J 1999 Generalization performance of support vector machines and other pattern classifiers *Advances in Kernel Methods: Support Vector Learning* (Cambridge, MA: MIT Press)
- [3] Bartlett P L and Mendelson S 2002 Rademacher and Gaussian complexities: risk bounds and structural results *J. Mach. Learn. Res.* **3** 463–82
- [4] Bartlett P L and Tewari A 2007 Sparseness vs estimating conditional probabilities: some asymptotic results *J. Mach. Learn. Res.* **8** 775–90
- [5] Bartlett P L, Long P M, Lugosi G and Alexander T 2020 Benign overfitting in linear regression *Proc. Natl Acad. Sci.* **117** 30063–70
- [6] Baum E B 1990 A polynomial time algorithm that learns two hidden unit nets *Neural Comput.* **2** 510–22
- [7] Belkin M, Hsu D and Xu J 2019 Two models of double descent for weak features (arXiv:1903.07571)
- [8] Boser B E, Guyon I M and Vapnik V N 1992 A training algorithm for optimal margin classifiers *Proc. 5th Annual Workshop on Computational Learning Theory* pp 144–52
- [9] Boufounos P T and Baraniuk R G 2008 One-bit compressive sensing *42nd Annual Conf. Information Sciences and Systems* (Piscataway, NJ: IEEE) pp 16–21
- [10] Buhot A and Gordon M B 2001 Robust learning and generalization with support vector machines *J. Phys. A: Math. Gen.* **34** 4377
- [11] Burges C J C 1996 Simplified support vector decision rules *Int. Conf. Machine Learning* vol 96 pp 71–7
- [12] Cao Y, Gu Q and Belkin M 2021 Risk bounds for over-parameterized maximum margin classification on sub-Gaussian mixtures *Advances in Neural Information Processing Systems* vol 34
- [13] Chatterji N S and Long P M 2020 Finite-sample analysis of interpolating linear classifiers in the overparameterized regime (arXiv:2004.12019)
- [14] Cortes C and Vapnik V 1995 Support-vector networks *Mach. Learn.* **20** 273–97
- [15] Cover T M 1965 Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition *IEEE Trans. Electron. Comput.* **EC-14** 326–34
- [16] Dietrich R, Oppen M and Sompolinsky H 1999 Statistical mechanics of support vector networks *Phys. Rev. Lett.* **82** 2975
- [17] Downs T, Gates K E and Masters A 2001 Exact simplification of support vector solutions *J. Mach. Learn. Res.* **2** 293–7
- [18] Gao W and Zhou Z-H 2013 On the doubt about margin explanation of boosting *Artif. Intell.* **203** 1–18
- [19] Germain P, Lacoste A, Laviolette F, Marchand M and Shanihan S 2011 A PAC-Bayes sample-compression approach to Kernel methods *ICML*
- [20] Graepel T, Herbrich R and Shawe-Taylor J 2005 PAC-Bayesian compression bounds on the prediction error of learning algorithms for classification *Mach. Learn.* **59** 55–76
- [21] Grønlund A, Kamma L and Larsen K G 2020 Near-tight margin-based generalization bounds for support vector machines (arXiv:2006.02175)

- [22] Hastie T, Montanari A, Rosset S and Tibshirani R J 2019 Surprises in high-dimensional ridgeless least squares interpolation (arXiv:1903.08560)
- [23] Keerthi S S, Chapelle O and DeCoste D 2006 Building support vector machines with reduced classifier complexity *J. Mach. Learn. Res.* **7** 1493–515
- [24] Klivans A R and Servedio R A 2008 Learning intersections of halfspaces with a margin *J. Comput. Syst. Sci.* **74** 35–48
- [25] Klivans A R, O'Donnell R and Servedio R A 2004 Learning intersections and thresholds of halfspaces *J. Comput. Syst. Sci.* **68** 808–40
- [26] Laurent B and Massart P 2000 Adaptive estimation of a quadratic functional by model selection *Ann. Stat.* **28** 1302–38
- [27] Liang T and Rakhlin A 2020 Just interpolate: kernel ‘ridgeless’ regression can generalize *Ann. Stat.* **48** 1329–47
- [28] Liu H 2019 Exact high-dimensional asymptotics for support vector machine (arXiv:1905.05125)
- [29] Mahdaviyeh Y and Naulet Z 2019 Risk of the least squares minimum norm estimator under the spike covariance model (arXiv:1912.13421)
- [30] Malzahn D and Oppen M 2005 A statistical physics approach for the analysis of machine learning algorithms on real data *J. Stat. Mech.* **P11001**
- [31] McAllester D 2003 Simplified PAC-Bayesian margin bounds *Learning Theory and Kernel Machines* (Berlin: Springer) pp 203–15
- [32] Mei S and Montanari A 2019 The generalization error of random features regression: precise asymptotics and double descent curve (arXiv:1908.05355)
- [33] Mitra P P 2019 Understanding overfitting peaks in generalization error: analytical risk curves for ℓ_2 and ℓ_1 penalized interpolation (arXiv:1906.03667)
- [34] Muthukumar V, Narang A, Subramanian V, Belkin M, Hsu D and Sahai A 2020 Classification vs regression in overparameterized regimes: does the loss function matter? (arXiv:2005.08054)
- [35] Muthukumar V, Vodrahalli K, Subramanian V and Sahai A 2020 Harmless interpolation of noisy data in regression *IEEE J. Sel. Areas Inf. Theory* **1** 67–83
- [36] Pisier G 1999 *The Volume of Convex Bodies and Banach Space Geometry* vol 94 (Cambridge: Cambridge University Press)
- [37] Rifkin R and Klautau A 2004 In defense of one-vs-all classification *J. Mach. Learn. Res.* **5** 101–41
- [38] Rudelson M and Vershynin R 2013 Hanson-wright inequality and sub-Gaussian concentration *Electron. Commun. Probab.* **18** 1–19
- [39] Schölkopf B and Smola A J 2002 *Learning with Kernels* (Cambridge, MA: MIT Press)
- [40] Schölkopf B, Smola A J, Williamson R C and Bartlett P L 2000 New support vector algorithms *Neural Comput.* **12** 1207–45
- [41] Steinwart I 2003 Sparseness of support vector machines *J. Mach. Learn. Res.* **4** 1071–105
- [42] Vapnik V N 1982 *Estimation of Dependences Based on Empirical Data* (Berlin: Springer)
- [43] Vapnik V N 1995 *The Nature of Statistical Learning Theory* (Berlin: Springer)
- [44] Vershynin R 2010 Introduction to the non-asymptotic analysis of random matrices (arXiv:1011.3027)
- [45] Wang K and Thrampoulidis C 2021 Benign overfitting in binary classification of Gaussian mixtures *ICASSP IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)* (Piscataway, NJ: IEEE) pp 4030–4
- [46] Wang K, Muthukumar V and Thrampoulidis C 2021 Benign overfitting in multiclass classification: all roads lead to interpolation *Advances in Neural Information Processing Systems* vol 34
- [47] Wang W and Fan J 2017 Asymptotics of empirical eigenstructure for high dimensional spiked covariance *Ann. Stat.* **45** 1342
- [48] Weston J and Watkins C 1998 Multi-class support vector machines *Technical Report* Citeseer
- [49] Zhang T 2002 Covering number bounds of certain regularized linear function classes *J. Mach. Learn. Res.* **2** 527–50