Journal of Machine Learning Research 22 (2021) 1-69 Submitted 6/20; Revised 5/21; Published 7/21

Classification vs regression in overparameterized regimes: Does the loss function matter?

Vidya Muthukumar* VMUTHUKUMAR8@GATECH.EDU
Electrical and Computer Engineering and Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA-30332, USA

Adhyyan Narang* ADHYYAN@UW.EDU
Department of Electrical and Computer Engineering
University of Washington
Seattle, WA-98115, USA

Vignesh Subramanian* VIGNESH.SUBRAMANIAN@EECS.BERKELEY.EDU
Department of Electrical Engineering and Computer Sciences
University of California Berkeley
Berkeley, CA-94720, USA

Mikhail Belkin MBELKIN@UCSD.EDU
Halıcıoğlu Data Science Institute
UC San Diego
La Jolla, CA-92093, USA

Daniel Hsu DJHSU@CS.COLUMBIA.EDU
Department of Computer Science and Data Science Institute
Columbia University
New York, NY-10027, USA

Anant Sahai SAHAI@EECS.BERKELEY.EDU
Department of Electrical Engineering and Computer Sciences
University of California Berkeley
Berkeley, CA-94720, USA

Editor: Simon Lacoste-Julien

Abstract

We compare classification and regression tasks in an overparameterized linear model with Gaussian features. On the one hand, we show that with sufficient overparameterization *all training points are support vectors*: solutions obtained by least-squares minimum-norm interpolation, typically used for regression, are identical to those produced by the hard-margin support vector machine (SVM) that minimizes the hinge loss, typically used for training classifiers. On the other hand, we show that there exist regimes where these interpolating solutions generalize well when evaluated by the 0-1 test loss function, but do not generalize if evaluated by the square loss function, i.e. they approach the null risk.

. *indicates equal contribution among authors who were students at the time of submission. Faculty are listed alphabetically after the students. The key results in this paper were unveiled at the ITA workshop in San Diego in February 2020.

Our results demonstrate the very different roles and properties of loss functions used at the training phase (optimization) and the testing phase (generalization).

Keywords: classification, regression, overparameterized, support vector machines, survival, contamination

1. Introduction

Paradigmatic problems in supervised machine learning (ML) involve predicting an output response from an input, based on patterns extracted from a (training) data set. In classification, the output response is (finitely) discrete and we need to classify input data into one of these discrete categories. In regression, the output is continuous, typically a real number or a vector. Owing to this important distinction in output response, the two tasks are typically treated differently. The differences in treatment manifest in two phases of modern ML: optimization (training), which consists of an algorithmic procedure to extract a predictor from the training data, typically by minimizing the training loss (also called *empirical risk*); and generalization (testing), which consists of an evaluation of the obtained predictor on a separate test, or validation, data set.

Traditionally, the choice of loss functions for both phases is starkly different across classification and regression tasks. The square loss function is typically used both for the training and the testing phases in regression. In contrast, the hinge or logistic (*cross-entropy* for multi-class problems) loss functions are typically used in the training phase of classification, even though the 0-1 loss function is used for testing. The use of the logistic and hinge losses can be motivated by their computational and statistical properties. For example, the theory of *surrogate losses* (Zhang, 2004; Bartlett et al., 2006; Ben-David et al., 2012) gives theoretical arguments favoring the logistic and hinge losses over other convex surrogates, including the square loss.[1] Yet, there have been indications that the reality is more complex, both in underparameterized and overparameterized regimes of ML. For example, Rifkin (2002) extensively compared the hard-margin *support vector machine* (SVM), which minimizes the hinge loss, and *regularized least-squares classification* (RLSC), which minimizes the square loss — ultimately concluding that "the performance of the RLSC is essentially equivalent to that of the SVM across a wide range of problems, and the choice between the two should be based on computational tractability considerations." Quite similar results[2] for a comparison between the square loss and cross-entropy loss have recently been obtained in Hui and Belkin (2021) for a range of modern neural architectures and data sets across several application domains. The latter is the current dominant standard for training neural networks.

In an important separate development, we have recently seen compelling evidence that overparameterized deep neural networks, as well as other models trained to *interpolate* the data (i.e. achieve zero, or near zero, training loss), are capable of good test performance, questioning conventional[3] statistical wisdom (Neyshabur et al., 2014; Zhang et al., 2016;

1. Also see, e.g., Section 8.1.2 in Goodfellow et al. (2016) for a representative informal discussion.
2. In fact, while the results were generally close, in a majority of classification tasks models trained using the square loss outperformed models trained with cross-entropy.
3. For example, Hastie, Tibshirani and Friedman say, on page 221 of their popular statistical learning textbook (Hastie et al., 2009), that "a model with zero training error is overfit to the training data and will typically generalize poorly."

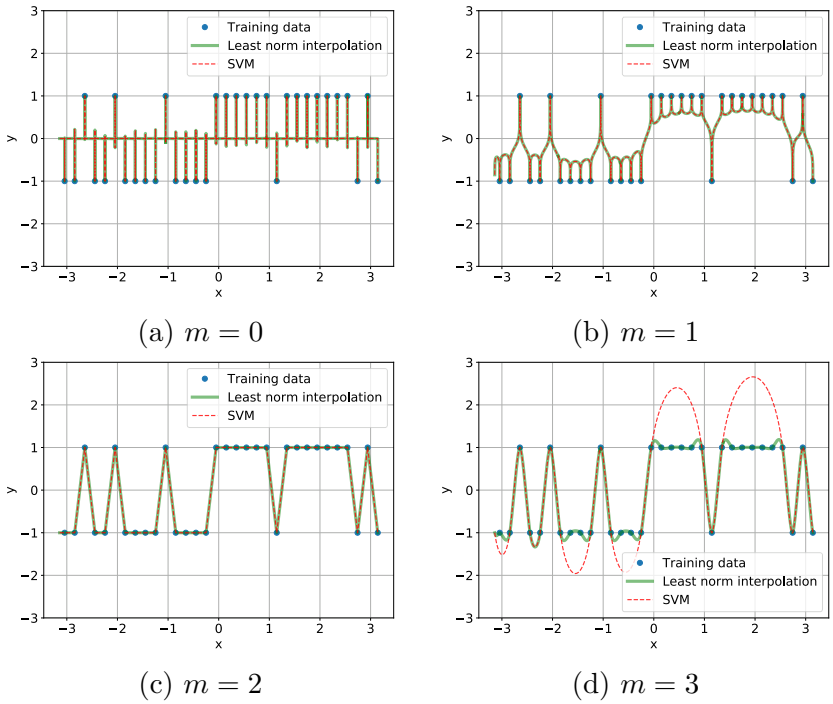
Belkin et al., 2018; Geiger et al., 2019; Belkin et al., 2019). Since then, the theoretical ML community has identified regimes *for regression tasks* under which overfitting is benign (to borrow the language from Bartlett et al., 2020), and interpolating *noisy* data with solutions arising from empirical risk minimization is compatible with good generalization (Bartlett et al., 2020; Belkin et al., 2020; Hastie et al., 2019; Mei and Montanari, 2019; Muthukumar et al., 2020). It is worth noting that the ensuing test loss of interpolating solutions is unrelated to their training loss, which is identically zero for all such solutions. This demonstrates, again, that the relationship between training and test losses — both for regression and classification tasks — is more complex than often assumed.

This paper introduces a direct comparison between the different loss functions used in classification and regression, in both the training and testing phases. We analyze the modern overparameterized regime under the linear model with Gaussian features and uncover a remarkable phenomenon of overparameterized training: in sufficiently overparameterized settings, with high probability, *every training data point is a support vector*. Consequently, the outcome of optimization (with gradient descent) is the same whether we use the hinge loss, logistic loss, or the square loss.

On the other hand, we show that the choice of test loss function results in a significant asymptotic difference between classification and regression tasks. In particular, we identify truly overparameterized regimes for which predictors will generalize poorly for regression tasks (measured by the square loss), but well for classification tasks (measured by the 0-1 loss). The fact that regression and classification can be different has been understood for some time. For example, Devroye, Gyorgi and Lugosi point out in Chapter 6.7 of their classic textbook (Devroye et al., 1991) that *"classification is easier than regression function estimation"*, in reference to sample complexity. Artificial examples are also given where accurate regression estimates of the density are impossible given the hypothesis class, but classification succeeds because of the separability of the two classes. In contrast, our paper demonstrates that in overparameterized regimes, this phenomenon can be quite generic. Approximability is not the underlying issue; rather, it is a consequence of learning.

1.1 Our contributions

Our study investigates differences and commonalities between classification and regression, using the overparameterized linear model with Gaussian features. On the side of commonality, we connect the hard-margin-maximizing SVM to the minimum-ℓ_2-norm interpolator of classification training data (i.e. binary labels), by showing that they are identical once the degree of "effective overparameterization" is sufficiently large (Theorem 11). This shows that, contrary to the prevailing low-dimensional intuition, there is no difference between maximum margin and least-squares solutions in high-dimensional settings. In particular, using the appropriate optimization methods, minimizing the logistic, hinge, and square loss yield exactly identical predictors. This phenomenon is not restricted to Gaussian feauturization: in fact, as Figure 1 illustrates, we first noticed that all training points tend to become support vectors in the case of Fourier features on regularly spaced training data. For a detailed description of this "ultra-toy" model and experiment, see Appendix A.



(a) $m = 0$ (b) $m = 1$

(c) $m = 2$ (d) $m = 3$

Figure 1: Correspondence between the SVM and minimum-ℓ_2-norm interpolation, illustrated by Fourier features on regularly spaced training data with 10% label noise (for various rates of feature scaling corresponding to $\lambda_k = \frac{1}{k^m}$ in the optimization to adjust the preference for lower frequencies as given in Definition 19 in Appendix A). For all the figures, the number of samples $n = 32$ and the number of features $d = 2^{14}$. Notice that below $m = 2$, all points are support vectors and the SVM solution agrees with least-norm interpolation.

In contrast, we show that the choice of loss function used on *test data* yields significant differences between classification and regression. Depending on the extent of "effective overparameterization", the same minimum-norm solution can:

- succeed at both regression and classification,
- succeed at classification and fail at regression, or
- fail at both,

as we show in Theorem 13. The intermediate regime of special interest is the one for which minimum-ℓ_2-norm interpolators generalize poorly in regression tasks, but well in classification tasks. Underlying these results is a sharp non-asymptotic analysis of the minimum-ℓ_2-norm interpolator *for the classification task*. We conceptually link the techniques introduced in recent analysis of this interpolator for the regression task (Bartlett et al., 2020) to the classification task, using a signal-processing (Fourier-theoretic) interpretation of the over-parameterized regime that was introduced in Muthukumar et al. (2020). This constitutes

a first analysis of this type for classification tasks, providing non-asymptotically matching upper and lower bounds. In Section 6, we demonstrate that the classic *upper bounds*, based on training data margin, fail to produce meaningful results or useful intuition in our setting.

2. Related work

The phenomenon of overparameterization and interpolation yielding significantly improved empirical performance across a variety of models as well as tasks (Neyshabur et al., 2014; Zhang et al., 2016; Geiger et al., 2019; Belkin et al., 2019) has received significant research attention over the last few years. In this section, we contextualize our results in this research landscape.

2.1 The role of the training loss function (and optimization algorithm)

At a high level, any solution obtained in an overparameterized regime that generalizes well must have some sort of *regularization*, i.e. special structural constraints on the values it can take. Thus, we need to understand the influence of the choice of training loss function on the resulting solution and its generalization guarantee. In the overparameterized regime, there are infinitely many solutions that interpolate training data, and indeed even more that *separate* discretely labeled data. Thus, characterizing the *implicit regularization* (Ji and Telgarsky, 2019; Soudry et al., 2018; Gunasekar et al., 2018; Woodworth et al., 2019; Nacson et al., 2019; Azizan et al., 2020) induced by the choice of optimization algorithm is important to understand properties of the obtained solutions. For the linear model, we have a concrete understanding of the solutions obtained by the most common choices of training loss functions:

1. If we minimize the logistic loss using gradient descent on separable training data[4], we will converge to the hard-margin SVM (Ji and Telgarsky, 2019; Soudry et al., 2018).

2. If we minimize the square loss on training data using gradient descent while also using an overparameterized model, we will converge to the minimum-ℓ_2-norm interpolation (Engl et al., 1996, Theorem 6.1) *provided the initialization is equal to zero.*

As mentioned in the introduction, conventional wisdom recommends the choice of the logistic loss, or the hinge loss, for classification tasks. It is sometimes implied (without theoretical justification) that instead minimizing the square loss would be suboptimal for generalization. However, our first main result (Theorem 11) shows that with sufficient overparameterization, the SVM itself interpolates the binary labels — as pictured in Figure 2, this implies an equivalence in solutions corresponding to several choices of training loss function. Moreover, our subsequent Theorem 13 shows that the *interpolating* solution generalizes well in classification tasks, for a wide range of overparameterized regimes. And when it does generalize poorly, so does the SVM! These results add theoretical weight to the empirical evidence (Rifkin, 2002; Que and Belkin, 2016) that the hinge loss (and, by

4. The implicit bias has also been characterized for the more difficult non-separable case (Ji and Telgarsky, 2019), but we focus here on separable training data as this will always be the case for an overparameterized setting.

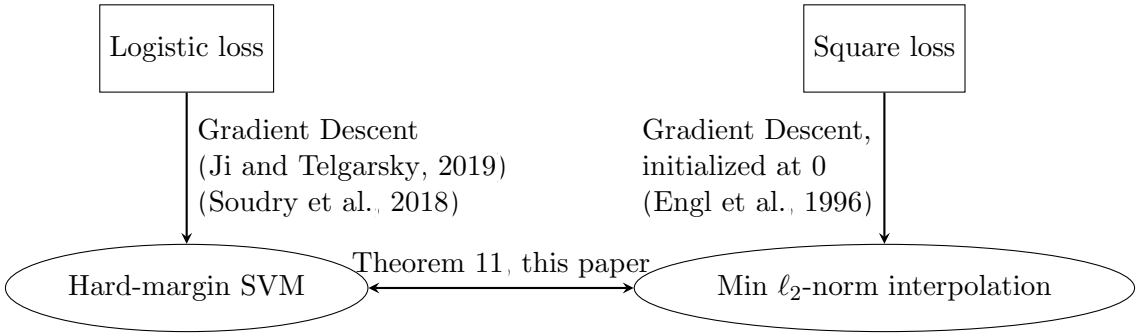


Figure 2: Equivalence of training procedures in overparameterized settings. Theorem 11 in this paper highlights exact equivalence *with high probability* between the hard-margin SVM and minimum-ℓ_2-norm interpolation under sufficient *effective overparameterization*.

extension, the cross-entropy loss) is not necessarily the superior choice for classification tasks.

Our Theorem 11 establishes a link between the hard-margin SVM and the minimum-ℓ_2-norm interpolation by exhibiting an overparameterized and separable setting where every training example is a support vector. Previous works have related the number of support vectors in *soft-margin* SVMs and the Bayes risk and conditional probability estimation (Steinwart, 2003; Bartlett and Tewari, 2007), but do not apply to the hard-margin SVM on separable data sets.

The SVM maximizes *training data margin* in feature space. Theoretical analyses of generalization error as a function of the margin have been proposed to explain the success of models such as boosting and neural networks (Schapire et al., 1998; Bartlett, 1998; Bartlett et al., 2017). Explanations based on margin bounds are sometimes credited, in a heuristic manner, for the empirical success of interpolated models in classification tasks. This is, in fact, a misleading explanation (as also noted in Shah et al., 2018); in Section 6, we provide experimental evidence for the tautology of generalization upper bounds as a function of the feature-space margin, when applied to sufficiently overparameterized models. This evidence corroborates the recent perspectives on modern ML which argue against generalization bounds that tie training loss to the expected loss on test data (Belkin et al., 2018; Nagarajan and Kolter, 2019). We instead favor a first-principles approach to analyzing high-dimensional models for classification, inspired by recent progress in regression.

2.2 Insights from least-squares regression

The recently observed phenomenon of *double descent* (Geiger et al., 2019; Belkin et al., 2019) made concrete explicit empirical benefits of overparameterization. Subsequent work (Bartlett et al., 2020; Belkin et al., 2020; Hastie et al., 2019; Mei and Montanari, 2019; Muthukumar et al., 2020; Mitra, 2019; Nakkiran, 2019) has identified theoretical conditions under which overparameterization and interpolation can be helpful, or at the very least, harmless in linear regression with different feature families. The main insight can be

crystallized as follows: for overparameterized solutions to interpolate "benignly", the feature family needs to satisfy a delicate balance between having a few important directions that favor the true signal (unknown function), and a large number of unimportant directions that *absorb* the noise in a harmless manner. This trade-off was explored both for minimum-ℓ_2-norm (Bartlett et al., 2020; Belkin et al., 2020; Hastie et al., 2019; Mei and Montanari, 2019; Muthukumar et al., 2020; Mitra, 2019) and minimum-ℓ_1-norm interpolations (Muthukumar et al., 2020; Mitra, 2019).

In this paper, we build on these insights from regression tasks, to show that overparameterized models can similarly generalize well for classification tasks. In fact, it turns out that the conditions for classification are milder, and there is an intermediate regime of overparameterization where the regression problem is "hard" — but the classification problem is "easy". Focusing on the well-specified case, we show that *the balance between preserving signal and absorbing noise does not have to be as delicate for classification tasks as it is for regression tasks*. This conclusion cannot be made directly from the regression analyses, as 0-1 classification error is quite different from the mean-square regression error. To bridge the gaps, we use a signal processing perspective on the overparameterized regime that was first developed in Muthukumar et al. (2020), where the conditions for low test error are linked to notions of *survival* of the true features and *contamination* by falsely discovered features. We will see (in Section 5.2) that these same quantities show up explicitly in the analysis of classification test error.

2.3 Recent work on high-dimensional classification/logistic regression

High-dimensional *logistic regression* and classification are naturally closely connected, and statisticians have studied the former in a number of contexts. Properties of penalized maximum likelihood estimators in overparameterized logistic regression have received substantial attention (an incomplete list is Bunea (2008); Van de Geer (2008); Kakade et al. (2010); Fan and Lv (2011)). Here, the penalty, or regularizer, is typically in ℓ_1-norm and its relatives, and the studied solutions do not interpolate training data. In contrast, our focus is on classification problems and thus the properties of the ℓ_2-margin-maximizing support-vector-machines, and moreover we make explicit connections to solutions that interpolate binary labels.

Most pertinent to our setting, we acknowledge a recent line of work (Deng et al., 2019; Montanari et al., 2019; Kini and Thrampoulidis, 2020) that identifies precise asymptotics for the generalization error of the SVM as a function of the overparameterization factor. The main technical tool common to these works is the convex Gaussian min-max theorem (Thrampoulidis et al., 2015), a generalization of Gordon's min-max theorem (Gordon, 1985) that has seen substantial application to obtain precise asymptotics in high-dimensional regression. It is worth noting that these elegant analyses specifically assume isotropic featurization, and do not study the ramifications of anisotropy, which is known to be critical for good generalization of ℓ_2-regularized solutions. While Kini and Thrampoulidis (2020) do compare the outcomes of logistic and square loss in binary classification, and show that the test error is almost identical under extreme overparameterization, they do not show an explicit link between the actual solutions.

Concurrent to our work, Chatterji and Long (2020) directly *upper bound* the generalization error of the max-margin SVM under overparameterized linear discriminant models that are not isotropic and, like us, explore how anisotropic the situation needs to be for good generalization. Both their results and techniques are quite different from ours in that they study the iterates of gradient descent leveraging the implicit regularization perspective of optimization algorithms.

3. Setup

We begin with some basic notation. Thereafter, we describe the setup for training and test data, evaluation of classification and regression tasks, and choices of featurization (in that order).

3.1 Basic notation

We describe basic notation for vectors, matrices, and functions.

3.1.1 Vector and matrix notation

Let $\mathbf{e}_i$ represent the i^{th} standard basis vector (with the dimension implicit). For a given vector $\mathbf{v}$, the functional $\mathsf{sgn}(\mathbf{v})$ denotes the sign operator applied element-wise. Let $\mu_i(\mathbf{M})$ denote the i^{th} largest eigenvalue of positive semidefinite matrix $\mathbf{M}$, and $\mu_{\mathsf{max}}(\mathbf{M})$ and $\mu_{\mathsf{min}}(\mathbf{M})$ denote in particular the maximal and minimal eigenvalue respectively. Further, we use $||\mathbf{M}||_{\mathsf{op}}, \mathsf{tr}(\mathbf{M})$ and $||\mathbf{M}||_{\mathsf{F}}$ to denote the operator norm, trace norm, and Frobenius norm respectively.

3.1.2 Function-specific notation

For two functions $f(n)$ and $g(n)$, we write $f \asymp g$ *iff* there exist universal positive constants (c, C, n_0) such that

$$c\,|g(n)| \leq |f(n)| \leq C\,|g(n)| \;\; \forall n \geq n_0.$$

(In most places where we apply the above inequality, the functions f and g are positive valued and so we automatically drop the absolute value signs.)

3.2 Data

Let $\mathcal{X}$ denote the space of input data. For classification, our training data are *input data-binary label* pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ taking values in $\mathcal{X} \times \{-1, +1\}$; for regression, the training data are *input data*-real output pairs $(X_1, Z_1), \ldots, (X_n, Z_n)$ taking values in $\mathcal{X} \times \mathbb{R}$. We assume that there is a feature map $\boldsymbol{\phi}\colon \mathcal{X} \to \mathbb{R}^d$, target linear function parameterized by $\boldsymbol{\alpha^*} \in \mathbb{R}^d$, and label noise parameter $0 \leq \nu^* < 1/2$ such that for every $i \in \{1, 2, \ldots, n\}$, we have

$$Z_i = \langle \boldsymbol{\phi}(X_i),\, \boldsymbol{\alpha^*} \rangle \text{ and} \tag{1}$$

$$Y_i = \begin{cases} \mathsf{sgn}(Z_i) \text{ with probability } (1 - \nu^*) \\ -\mathsf{sgn}(Z_i) \text{ with probability } \nu^*. \end{cases} \tag{2}$$

Here, the feature map $\boldsymbol{\phi}$ is known, but the target parameter $\boldsymbol{\alpha}^*$ (which we refer to as the signal) is unknown. The label noise in Y_i is assumed to be independent of everything else.

Let $\boldsymbol{\phi}(x) = \begin{bmatrix}\phi_1(x) & \dots & \phi_d(x)\end{bmatrix}^T$ for $x \in \mathcal{X}$, i.e. $\phi_j(x)$ is the value of the j^{th} feature in $\boldsymbol{\phi}(x)$. We will consider the training data $\{X_i\}_{i=1}^n$ to be mutually independent and identically distributed (iid). Let $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{\phi}(X)\boldsymbol{\phi}(X)^\top]$ denote the covariance matrix of the feature vector $\boldsymbol{\phi}(X)$ for X following the same distribution as X_i. We assume $\boldsymbol{\Sigma}$ is invertible, so its square-root-inverse $\boldsymbol{\Sigma}^{-1/2}$ exists.

We define shorthand notation for the training data: let

$$\boldsymbol{\Phi}_{\mathsf{train}} := \begin{bmatrix}\boldsymbol{\phi}(X_1) & \boldsymbol{\phi}(X_2) & \cdots & \boldsymbol{\phi}(X_n)\end{bmatrix}^\top \in \mathbb{R}^{n\times d}$$

denote the data (feature) matrix; $\mathbf{Z}_{\mathsf{train}} := \begin{bmatrix}Z_1 & \dots & Z_n\end{bmatrix}^\top \in \mathbb{R}^n$ denote the regression output vector; and $\mathbf{Y}_{\mathsf{train}} := \begin{bmatrix}Y_1 & \dots & Y_n\end{bmatrix}^\top$ denote the classification output vector. Note that if there is no label noise (i.e. $\nu^* = 0$), then we have $\mathbf{Y}_{\mathsf{train}} = \mathsf{sgn}(\mathbf{Z}_{\mathsf{train}})$.

3.3 Classification, regression, and interpolation

The overparameterized regime constitutes the case in which the dimension (or number) of features is greater than the number of samples, i.e. $d \geq n$. We define the two types of solutions that we will primarily consider in this regime, starting with interpolating solutions.

Definition 1 *We consider solutions $\boldsymbol{\alpha}$ that satisfy one of the following feasibility conditions for interpolation:*

$$\boldsymbol{\Phi}_{\mathsf{train}}\boldsymbol{\alpha} = \mathbf{Y}_{\mathsf{train}} \text{ or} \tag{3a}$$

$$\boldsymbol{\Phi}_{\mathsf{train}}\boldsymbol{\alpha} = \mathbf{Z}_{\mathsf{train}} \tag{3b}$$

In particular, we denote the minimum-ℓ_2-norm interpolation on binary labels as

$$\widehat{\boldsymbol{\alpha}}_{2,\mathsf{binary}} := \operatorname*{arg\,min}_{\boldsymbol{\alpha}\in\mathbb{R}^d} \|\boldsymbol{\alpha}\|_2 \text{ s.t. Equation (3a) holds.}$$

Similarly, we denote the minimum-ℓ_2-norm interpolation on real labels as

$$\widehat{\boldsymbol{\alpha}}_{2,\mathsf{real}} := \operatorname*{arg\,min}_{\boldsymbol{\alpha}\in\mathbb{R}^d} \|\boldsymbol{\alpha}\|_2 \text{ s.t. Equation (3b) holds.}$$

Recall from our discussion in Section 2.1 that these interpolations arise from minimizing the square loss on training data. If we instead minimized the logistic or hinge loss, we would obtain the hard-margin *support vector machine* (SVM), defined below.

Definition 2 *For linearly separable data, the hard-margin Support Vector Machine (SVM) is $\widehat{\boldsymbol{\alpha}}_{\mathsf{SVM}} \in \mathbb{R}^d$, defined by*

$$\begin{aligned}\widehat{\boldsymbol{\alpha}}_{\mathsf{SVM}} := \operatorname*{arg\,min}_{\boldsymbol{\alpha}\in\mathbb{R}^d} \quad & \|\boldsymbol{\alpha}\|_2 \\ \text{s.t.} \quad & Y_i\boldsymbol{\phi}(X_i)^\top\boldsymbol{\alpha} \geq 1 \quad \text{for all } i = 1,\dots,n.\end{aligned} \tag{4}$$

Note that data is defined to be linearly separable iff the constraints in Equation (4) can be feasibly satisfied by some parameter vector $\boldsymbol{\alpha}$.

As long as $d \geq n$, any solution that interpolates the binary labels $\{Y_i\}_{i=1}^n$ satisfies Equation (4) with equality *almost surely* for any continuous distribution on the features. Thus, in the overparameterized regime, the training data is trivially linearly separable. Note, however, that the feasibility constraints do not require the SVM solution to interpolate the binary labels.

The standard metrics for test error in regression and classification tasks are, respectively, the mean-square-error (MSE) and classification error, defined as follows. In these definitions, we have ignored the irreducible error terms arising from possible additive noise in real outputs and label noise in binary outputs respectively. This reflects the practical goal of all prediction to get the underlying true output right, as opposed to matching noisy measurements of that underlying true output.

Definition 3 *The excess mean-square-error (MSE) of $\widehat{\boldsymbol{\alpha}} \in \mathbb{R}^d$ is*

$$\mathcal{R}(\widehat{\boldsymbol{\alpha}}) := \mathbb{E}[\langle \boldsymbol{\phi}(X),\, \boldsymbol{\alpha}^* - \widehat{\boldsymbol{\alpha}} \rangle^2]. \tag{5}$$

The excess classification error of $\widehat{\boldsymbol{\alpha}} \in \mathbb{R}^d$ is given by

$$\begin{aligned} \mathcal{C}(\widehat{\boldsymbol{\alpha}}) &:= \mathbb{E}\left[\mathbb{I}\left[\mathsf{sgn}(\langle \boldsymbol{\phi}(X),\, \boldsymbol{\alpha}^* \rangle) \neq \mathsf{sgn}(\langle \boldsymbol{\phi}(X),\, \widehat{\boldsymbol{\alpha}} \rangle)\right]\right] \\ &= \Pr\left[\mathsf{sgn}(\langle \boldsymbol{\phi}(X),\, \boldsymbol{\alpha}^* \rangle) \neq \mathsf{sgn}(\langle \boldsymbol{\phi}(X),\, \widehat{\boldsymbol{\alpha}} \rangle)\right]. \end{aligned} \tag{6}$$

Here, all expectations (and ensuing probabilities) are only over the random sample X of test data. As is standard, we will characterize the regression and classification test errors with high probability over the randomness in the training data $\{X_i, Y_i\}_{i=1}^n$.

As a final comment, we will typically construct an empirical estimate of both test error metrics from n_{test} test samples of data drawn without any label noise. This is for ease of empirical evaluation.

3.4 Featurization

We consider zero-mean Gaussian featurization, i.e. for every $i \in \{1, \ldots, n\}$, we have

$$\phi(X_i) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}). \tag{7}$$

We denote the spectrum of the (positive definite) covariance matrix $\boldsymbol{\Sigma}$ by the vector $\boldsymbol{\lambda} := \begin{bmatrix} \lambda_1 & \ldots & \lambda_d \end{bmatrix}$, where the eigenvalues are sorted in descending order, i.e. we have $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d > 0$.

Throughout, we will consider various *overparameterized ensembles* obtained by scaling the covariance parameter $\boldsymbol{\Sigma}$ as a function of both the number of training data points, n, and the number of features, d. We theoretically characterize the performance of solutions for classification and regression tasks using two representative ensembles, defined below.

Definition 4 (Isotropic ensemble(n, d)) *The isotropic ensemble, parameterized by (n, d), considers isotropic Gaussian features, $\boldsymbol{\Sigma} = \mathbf{I}_d$. For this ensemble, we will fix n and study the evolution of various quantities as a function of $d \geq n$.*

Note that the isotropic ensemble constitutes the "maximal" level of effective overparameterization (as defined in the second effective rank in Bartlett et al. (2020)) for a given choice of (n, d).

Definition 5 (Bi-level ensemble(n, p, q, r)) *The bi-level ensemble is parameterized by (n, p, q, r), where[5] $p > 1, 0 \leq r < 1$ and $0 < q < (p - r)$. Here, parameter p controls the extent of artificial overparameterization), r sets the number of preferred features, and q controls the weights on preferred features and thus effective overparameterization. In particular, this ensemble sets parameters*

$$\begin{aligned} d &:= n^p \\ s &= n^r \text{ and} \\ a &= n^{-q}. \end{aligned}$$

The covariance matrix of the Gaussian features $\mathbf{\Sigma}(p, q, r)$ is set to be a diagonal matrix, whose entries are given by:

$$\lambda_j = \begin{cases} \frac{ad}{s}, & 1 \leq j \leq s \\ \frac{(1-a)d}{d-s}, & \text{otherwise.} \end{cases}$$

For this ensemble, we will fix (p, q, r) and study the evolution of various quantities as a function of n.

The bi-level covariance matrix is parameterized by the choice for the top s eigenvalues and the bottom $(d-s)$ eigenvalues, with the sum of eigenvalues being invariant(equal to d). The parameters of critical importance are p, which determines the extent of overparameterization (i.e. number of features), r, which determines the number of larger eigenvalues, and q, which determines the relative values of larger and smaller eigenvalues (all as a function of the number of training points n). We make a few remarks below on this ensemble.

Remark 6 *This bi-level ensemble is inspired by the study of estimation of high-dimensional spiked covariance matrices (e.g. Wang and Fan, 2017; Mahdaviyeh and Naulet, 2019) when the number of samples is much smaller than the dimension. In these spiked matrices, the parameter s is typically set to a constant (that does not grow with n), and the top s eigenvalues are highly spiked with respect to the other $(d - s)$ eigenvalues. In fact, it is assumed that there exists a universal positive constant C, such that the smaller eigenvalues are bounded and the top (larger) eigenvalues grow with (d, n) in the following way:*

$$\lambda_j \geq \frac{d}{Cn} \quad \text{for all } j \in \{1, \ldots, s\} \tag{8a}$$

$$\lambda_j \leq C \quad \text{for all } j \in \{s+1, \ldots, d\}. \tag{8b}$$

Under these conditions, the ratio of the top to the bottom eigenvalues grows as $\Omega\left(\frac{d}{n}\right)$, and Wang and Fan (Wang and Fan, 2017) show that the top s estimated eigenvalues of the high-dimensional covariance matrix can be estimated reliably from samples, even when less than

5. We restrict (p, q, r) to this range to ensure that a) the regime is truly overparameterized (choice of p), b) the eigenvalues of the ensuing covariance matrix are always positive and ordered correctly (choice of q), c) the number of "high-energy" directions is sub-linear in n (choice of r).

the dimension (i.e. $n < d$). This condition, which is also critical for good generalization[6] *in regression problems, can be verified to be equivalent to the condition $q \leq (1 - r)$ in our bi-level ensemble (see Theorem 13 for a full statement). Our definition of the bi-level ensemble allows further flexibility in the choice of these parameters, and we will later show that classification tasks can generalize well even in the absence of this condition.*

Remark 7 *The bi-level ensemble can be verified to match the isotropic ensemble (Definition 4) as a special case when the parameters are set as $q + r = p$. This case represents the maximal level of effective overparameterization, and in general we take $q \leq (p - r)$ to ensure correct ordering of the eigenvalues. The smaller the value of q, the less the effective overparameterization. The models of Chatterji and Long (2020) are spiritually related in how they also use an exponent like q to control the effective overparameterization.*

Remark 8 *We know that for "benign overfitting" (Bartlett et al., 2020) of additive noise to occur in regression problems, we need to have sufficiently many (growing super-linearly in n) "unimportant" directions, corresponding to the lower level of eigenvalues. The choice of parameters $p > 1$ and $r < 1$ ensures that the number of such "unimportant" directions is equal to $(d - s) = (n^p - n^r) \gg n$, and so the bi-level ensemble as defined does not admit the regime of harmful overfitting of noise for any choice of parameters (p, q, r). This allows us to isolate signal shrinkage as the principal reason for large regression error, and also study the ramifications of such shrinkage for classification error.*

In addition to the above, we empirically study (in Section 6) the behavior of various quantities for two other ensembles defined below, both of which have been previously studied in regression tasks.

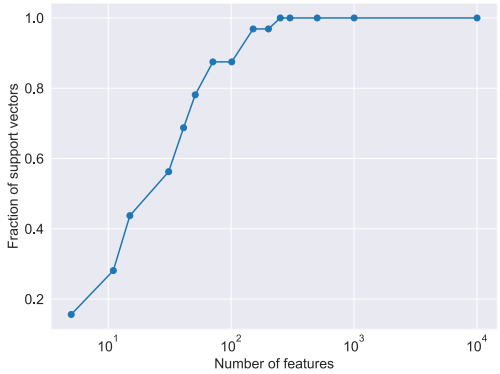
Definition 9 ("Weak features" ensemble) *This ensemble is a simplification of feature families introduced by Belkin et al. (2020); Hastie et al. (2019); Mei and Montanari (2019), all of which demonstrate an explicit benefit of overparameterization in generalization for regression tasks. The features consist of raw, 1-dimensional random variable $X_i \sim \mathcal{N}(0, \sigma^2)$, and independent d-dimensional random variables $\mathbf{W}_i$ i.i.d. $\sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ for $i \in \{1, \ldots, n\}$. Then, for a given value of $d > n$, each lifted feature is given by:*

$$\boldsymbol{\phi}(X_i) = X_i \mathbf{1_d} + \mathbf{W}_i. \tag{9}$$

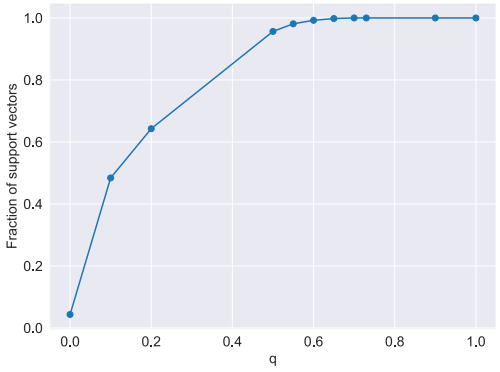
We define the real output Z_i, corresponding to input X_i, as $Z_i = X_i$, and similarly the binary output Y_i is defined as $Y_i = \mathsf{sgn}(X_i)$. For this ensemble, we will fix n and study the evolution of various quantities as a function of d. This model allows us to study model mis-specification because the real output is not exactly representable in the feature space.

Define the column vector of raw features as $\mathbf{X_{train}} = \begin{bmatrix} X_1 & X_2 & \ldots & X_n \end{bmatrix}^\top$ and the matrix $\mathbf{W} \in \mathbb{R}^{n \times d}$ whose i_{th} row is $\mathbf{W}_i$. For the above featurization, the training matrix admits the simple form $\begin{bmatrix} \mathbf{X_{train}} & \mathbf{X_{train}} & \ldots & \mathbf{X_{train}} \end{bmatrix} + \mathbf{W}$.

6. In particular, avoiding signal shrinkage, as also shown in Bartlett et al. (2020).



(a) Isotropic features with $n = 32$ fixed. Equation (12) would require $d \geq 1141$.



(b) Bi-level ensemble with $(n = 529, d = 12167)$ fixed and parameters $(p = 3/2, r = 1/2)$ fixed. As the parameter q increases, the fraction of support vectors increases.

Figure 3: Experimental illustration of Theorem 11 for Gaussian features: fraction of the training points that are support vectors increases as effective overparameterization increases.

Definition 10 ("Polynomial decay of eigenvalues" ensemble) *This ensemble is inspired by commonly chosen reproducing kernel Hilbert spaces and is parameterized by $m \geq 0$. We set the spectrum of the covariance matrix $\mathbf{\Sigma}$ to be*

$$\lambda_k = \frac{1}{k^m} \text{ for all } k \in \{1, 2, \ldots, d\}. \tag{10}$$

For this ensemble, we will fix (n, d) and study the evolution of various quantities as a function of the parameter m.

4. Approximating the SVM by exact interpolation

From the optimization objective and constraints defined in Equation (4), we can see that there is a continuum of margins, defined by $Y_i \cdot \phi(X_i)^\top \alpha$, that is possible for each training point. Thus, unlike in least-squares regression, even obtaining an exact expression for the margin-maximizing SVM solution, $\widehat{\boldsymbol{\alpha}}_{\mathsf{SVM}}$, appears difficult in the overparameterized regime.

The heart of our approach to tractably analyzing the SVM involves making an explicit link to minimum-ℓ_2-norm interpolation, by showing that *all the training data points usually become support vectors* in a sufficiently overparameterized regime. We actually first identified this phenomenon in visualizations of the SVM and the minimum-ℓ_2-norm interpolation using *Fourier features* on one-dimensional data; details of this auxiliary experiment are contained in Appendix A.1. The number of support vectors was also recently empirically

observed[7] to increase with the number of model parameters (Snyder and Vishwanath, 2020). In fact, we believe that these ideas are spiritually connected to the well-known folk wisdom that the number of support vectors tends to proliferate when increasing the "bandwidth" of kernels like the radial basis function (RBF) kernel.

Such a phenomenon, when true, explicitly links the concept of support-vector-machines with a positive margin constraint to exact interpolation of the training data labels, and suggests a roadmap to analyzing the generalization error of the SVM by analyzing the latter solution (which we do subsequently in Section 5). We show in the theorem below that this phenomenon manifests with high probability provided there is sufficient *effective overparameterization*.

Theorem 11 *Let $\mathbf{\Phi}_{\mathsf{train}}$ follow the Gaussian featurization from Equation* (7) *with covariance matrix $\mathbf{\Sigma}$, and let $\widehat{\boldsymbol{\alpha}}_{\mathsf{SVM}}$ be the solution to the optimization problem in Equation* (4)*.*

1. *If $\mathbf{\Sigma}$ satisfies*

$$||\boldsymbol{\lambda}||_1 \geq 72 \left(||\boldsymbol{\lambda}||_2 \cdot n\sqrt{\ln n} + ||\boldsymbol{\lambda}||_\infty \cdot n\sqrt{n} \ln n + 1 \right), \tag{11}$$

   the vector $\widehat{\boldsymbol{\alpha}}_{\mathsf{SVM}}$ satisfies the binary label interpolation constraint (Equation (3a)*) simultaneously for every $\mathbf{Y}_{\mathsf{train}} \in \{\pm 1\}^n$ with probability at least $\left(1 - \frac{2}{n}\right)$.*

2. *If $\mathbf{\Sigma} = \mathbf{I}_d$ (i.e., $\mathbf{\Phi}_{\mathsf{train}}$ follows the isotropic ensemble), and*

$$d > 10n \ln n + n - 1, \tag{12}$$

   then the vector $\widehat{\boldsymbol{\alpha}}_{\mathsf{SVM}}$ satisfies the binary label interpolation constraint (Equation (3a)*) for any fixed $\mathbf{Y}_{\mathsf{train}} \in \{\pm 1\}^n$ with probability at least $\left(1 - \frac{2}{n}\right)$.*

Theorem 11 is proved in Appendix C. Both conditions are proved by showing that a complementary slackness condition on the dual of the SVM optimization problem holds with high probability — where the conditions differ is in the application of concentration bounds. The condition in Equation (11) is proved using a broadly applicable "epsilon-net" argument to bound the operator norm of a random matrix, while the sharper condition in Equation (12) leverages Gaussian isotropy and precise properties of the inverse Wishart distribution. Note that the first result holds for all label vectors $\mathbf{Y}_{\mathsf{train}} \in \{\pm 1\}^n$ simultaneously, while the second result holds for any fixed $\mathbf{Y}_{\mathsf{train}}$ (but independent of the features).

We now remark on the result for the isotropic and the bi-level ensemble.

Remark 12 *Plugging in the condition in Equation* (11) *into the bi-level ensemble (Definition 5), the following conditions on (p, q, r) are sufficient for all training points to become support vectors with high probability (see Appendix C.2 for a full calculation):*

$$p > 2 \text{ and} \tag{13a}$$

$$q > \left(\frac{3}{2} - r \right). \tag{13b}$$

7. In (Snyder and Vishwanath, 2020), the interest is primarily in showing that the number of support vectors constrains the complexity of the learned model and can be related to generalization performance. Our results here show that good generalization is possible even when everything becomes a support vector.

There is an intuitive interpretation for each of these conditions in light of the second "effective rank" condition that is sufficient for benign overfitting (Bartlett et al., 2020) of noise (although our proof technique is quite different). First, the condition $p > 2$ mandates an excessively large number of unimportant directions, i.e. corresponding to lower-level (smaller) eigenvalues ($(n^p - n^r)$ of them). Second, the condition $q > \left(\frac{3}{2} - r\right)$ mandates that the ratio between the important directions, i.e. higher-level eigenvalues, and the unimportant directions, is sufficiently small — thus, the unimportant directions are sufficiently weighted. This second condition appears to be strictly stronger than what is required for benign overfitting of noise.

Equation (13) *is quite strong as a sufficient condition, but nevertheless admits non-trivial regimes for which classification can generalize well or poorly (see the text accompanying Theorem 13 for a full discussion). However, there is ample evidence to suggest that this condition is not necessary. Notably, Figure 3(b) shows that with a choice of parameterization* $(p = 3/2, r = 1/2)$*, the fraction of support vectors becomes equal to* 1 *around when* $q \geq 0.7$*. This choice of parameters for the bi-level ensemble clearly violates both conditions in Equation* (13)*. Thus, all training points become support vectors more often than our theory predicts. Subsequent work to ours (Hsu et al., 2021) tightened the condition in Equation* (11) *by providing a new deterministic equivalent to the phenomenon of all training points becoming support vectors.*

From this section, we have identified an interesting phenomenon by which all training points become support vectors with sufficient effective overparameterization. Moreover, this phenomenon is even more prevalent empirically than our current theory predicts.

5. Generalization analysis for interpolating solution with Gaussian features

In Section 4, we showed that the SVM solution often exactly corresponds to the *minimum-ℓ_2-norm* interpolation on binary labels, denoted by $\widehat{\boldsymbol{\alpha}}_{2,\mathsf{binary}}$. In this section, we attempt an approximate characterization of the ensuing classification error of this interpolation. Our hope is that we can leverage comprehensive analyses of minimum-ℓ_2-norm interpolation for least-squares regression (Bartlett et al., 2020; Muthukumar et al., 2020). However, it turns out that direct plug-ins of these analyses do not work for a number of reasons:

1. Even with clean data (i.e. zero label noise), the classification setup admits misspecification noise of the form $Y_i - \boldsymbol{\phi}(X_i)^\top \boldsymbol{\alpha}^*$. The misspecification noise is clearly non-zero mean, and is non-trivially correlated with the features. This resists a clean decomposition of generalization error into the error arising from signal identifiability (or lack thereof) + error arising from overfitting of noise, as in Bartlett et al. (2020).

2. For a given interpolation $\widehat{\boldsymbol{\alpha}}$, the expression for classification error is distinctly different from mean-square-error (we will see this explicitly in Theorem 17). In particular, we will see that characterizing this expression sharply requires novel analysis of the individual recovered coefficients as a result of interpolation.

Our analysis is subsequently non-trivial to engage with both of these difficulties, and directly addresses both of them by analyzing the minimum-ℓ_2-norm interpolator of binary labels

from first principles. This is, roughly speaking, in two steps: first, by characterizing the expected generalization error in terms of 0-1 classification loss *for any solution* (regardless of whether it interpolates or not) as a function of *survival* and *contamination* factors; second, by obtaining sharp characterizations of these factors for the minimum-ℓ_2-norm interpolator of binary labels.

5.1 Setup and result

We state our main result for this section in the context of the bi-level ensemble (Definition 5). We fix parameters $p > 1$ (which represents the extent of artificial overparameterization), and $r \in [0, 1)$ (which sets the number of preferred features), and $q \in [0, p - r]$ (which controls the weights on preferred features, thus effective overparameterization); and study the evolution of regression and classification risk as a function of n. For the purpose of this section, we denote the regression and classification test losses under the bi-level ensemble as $\mathcal{R}(\widehat{\boldsymbol{\alpha}}_{2,\mathsf{real}}; n)$ and $\mathcal{C}(\widehat{\boldsymbol{\alpha}}_{2,\mathsf{binary}}; n)$, to emphasize that these losses vary with n.

In addition to this and the broad setup as described in Section 3 we make a 1-*sparse assumption* on the unknown parameter vector $\boldsymbol{\alpha}^*$, as described below.

Assumption 1 (1-sparse linear model) *Recall that the bi-level ensemble sets $s := n^r$. For some unknown*[8] *$t \in \{1, \ldots, s\}$, we assume that $\boldsymbol{\alpha}^* = \frac{1}{\sqrt{\lambda_t}} \cdot \mathbf{e}_t$, i.e. the parameter vector $\boldsymbol{\alpha}^*$ is 1-sparse.*

Assumption 1 is most useful to for us to derive clean expressions for regression and classification error in terms of natural notions of "survival" and "contamination", as detailed subsequently in Section 5.2. While this assumption appears rather strong, it is actually without loss of generality within the bi-level ensemble *for analyzing the performance of minimum-ℓ_2-norm interpolation specifically.* If the true parameter vector $\boldsymbol{\alpha}^*$ has support only within the s favored directions, then we can choose another orthonormal coordinate system in which this $\boldsymbol{\alpha}^*$ is only along the first direction. Because minimum-ℓ_2-norm interpolation does not care about orthonormal coordinate changes and such a change will not change the underlying covariance matrix, we just assume 1-sparsity to capture the representability of the true model by the favored features.

Under Assumption 1, we now show the existence of a regime, corresponding to choice of (p, q, r) above, for which the regression test loss stays prohibitively high, but the classification test loss goes to 0 as $n \to \infty$. (We also derive non-asymptotic versions of these results in Appendix E, but only state the asymptotic results here for brevity.)

Theorem 13 *Assume that the true data generating process is 1-sparse (Assumption 1). For the bi-level covariance matrix model, the limiting classification and regression error of the minimum-ℓ_2-norm interpolation (of binary labels and real labels respectively) converge in probability, over the randomness in the training data, as a function of the parameters (p, q, r) in the following way:*

8. The intuition for this condition, also motivated in prior analyses of minimum-ℓ_2-norm interpolation (Muthukumar et al., 2020), is that for any reasonable preservation of signal, the true feature needs to be sufficiently preferred, therefore weighted highly.

1. *For* $0 \leq q < (1-r)$*, we have*

$$\lim_{n\to\infty} \mathcal{R}(\widehat{\boldsymbol{\alpha}}_{2,\mathsf{real}}; n) = 0,$$
$$\lim_{n\to\infty} \mathcal{C}(\widehat{\boldsymbol{\alpha}}_{2,\mathsf{binary}}; n) = 0.$$

   In this regime, both regression and classification generalize well.

2. *For* $(1-r) < q < (1-r) + \frac{(p-1)}{2}$*, we have*

$$\lim_{n\to\infty} \mathcal{R}(\widehat{\boldsymbol{\alpha}}_{2,\mathsf{real}}; n) = 1,$$
$$\lim_{n\to\infty} \mathcal{C}(\widehat{\boldsymbol{\alpha}}_{2,\mathsf{binary}}; n) = 0.$$

   In this regime, classification generalizes well but regression does not.

3. *For* $(1-r) + \frac{(p-1)}{2} < q \leq (p-r)$*, we have*

$$\lim_{n\to\infty} \mathcal{R}(\widehat{\boldsymbol{\alpha}}_{2,\mathsf{real}}; n) = 1,$$
$$\lim_{n\to\infty} \mathcal{C}(\widehat{\boldsymbol{\alpha}}_{2,\mathsf{binary}}; n) = \frac{1}{2}.$$

   In this regime, the generalization is poor for both classification and regression.

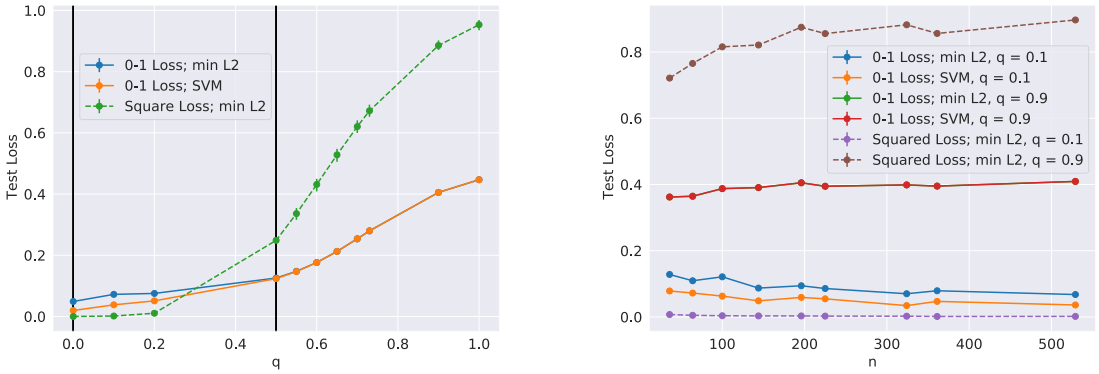
Note that the presence of label noise ν^* *does not affect these asymptotic scalings (since* $\nu^* < 0.5$*).*

Figure 4(a) shows the evolution of classification and regression error as a function of the parameter q, fixing $p = 3/2$ and $r = 1/2$. The classification error is plotted for both the SVM and the minimum-ℓ_2-norm interpolation — as we expect from Theorem 11, these are remarkably similar. Figure 4(b) shows that the empirical quantities converge to the limiting quantities from Theorem 13.

The new regime of principal interest that we have identified is values of $q \in (1-r, 1-r+\frac{p-1}{2})$ for which classification generalizes, but regression does not. The entire proof of Theorem 13 is deferred to Appendices D and E, but we briefly illustrate the intuition for this discrepancy between classification and regression tasks in Section 5.2. In particular, we will see that good generalization for classification requires a far less stringent condition on coefficient recovery than regression.

We now provide some intuition for the scalings described in Theorem 13 for the bi-level ensemble.

Remark 14 *Observe that in this ensemble, regression tasks generalize well iff we have* $q < (1-r)$*, which is a condition directly related to signal preservation. Recall that for fixed values of* (p, r)*, the parameter* q *controls the relative ratio of the larger eigenvalues to the smaller eigenvalues (corresponding to unimportant directions). The higher the value of* q*, the smaller this ratio, and the harder it is to preserve signal. The results on "benign overfitting" (Bartlett et al., 2020) upper bound the contribution of (bounded* ℓ_2*-norm) pure signal to regression error. This upper bound can also be verified to decay with* n *iff we have* $q \leq (1-r)$*. Furthermore, as we already remarked on Definition 5, the bi-level ensemble is designed to always avoid harmful noise overfitting. (We will, however, see in the next remark that the rate of effective noise absorption is important.)*



(a) Test loss vs q

(b) Test loss vs number of training points

Figure 4: Comparison of test classification and regression error on solutions obtained by minimizing different choices of training loss on the bi-level ensemble. For both figures, parameters $(p = 3/2, r = 1/2)$ are fixed. On the left, $n = 529, d = 12167$ are fixed. Here, the dashed green curve corresponds to $\widehat{\boldsymbol{\alpha}}_{2,\mathsf{real}}$ (Equation 3b), the orange curve corresponds to $\widehat{\boldsymbol{\alpha}}_{2,\mathsf{binary}}$ (Equation 3a), the solid blue curve corresponds to $\widehat{\boldsymbol{\alpha}}_{\mathsf{SVM}}$ (Equation 4), and the black lines demarcate the regimes from Theorem 13. On the right, d varies as $n^{\frac{3}{2}}$.

Remark 15 *The regime that we have identified that is of principal interest is intermediate values of q, i.e. $(1-r) < q < (1-r) + \frac{(p-1)}{2}$. This highlights a fascinating role that overparameterization, in the form of the parameter p, plays in allowing the good generalization of interpolating solutions in classification tasks. Recall that the larger the value of p, the larger the total number of features $d = n^p$. Thus, there are several "unimportant directions" in the bi-level ensemble all corresponding to the smaller eigenvalue — which helps in harmless absorption of effective noise. In the proof of Theorem 13, we will identify an explicit mechanism by which having many unimportant directions helps in good generalization for classification, even though the signal is not preserved. At a high level, this mechanism constitutes the spreading out of attenuated signal across several features in a relatively "harmless" way, to exhibit minimal influence on classification performance. In fact, this influence is quantified by a notion of "contamination" by falsely discovered features (defined in Section 5.2) that can be directly linked to the contribution of noise overfitting to regression error.*

Finally, we remark that Theorem 13 provides a connection between classification and regression test error when both tasks are solved using the minimum-ℓ_2-norm interpolation, i.e. minimizing the square loss on training data. Since we explicitly linked the minimum-ℓ_2-norm interpolation and the SVM in the preceding Section 4, it is natural to ask whether the generalization results in Theorem 13 help us directly compare the SVM for classification tasks and the minimum-ℓ_2-norm interpolation for regression tasks. We can indeed do this in a slightly more restricted regime of the bi-level ensemble, described below.

Corollary 16 *Assume that the true data generating process is* 1*-sparse (Assumption 1). Consider the bi-level ensemble with* $p > 2$*. Then, the classification error of the SVM (on binary labels), and the regression error of the minimum-*ℓ_2*-norm interpolation (on real labels), converge in probability as follows:*

1. *For* $\left(\frac{3}{2} - r\right) < q < (1 - r) + \frac{(p-1)}{2}$*, we have*

$$\lim_{n\to\infty} \mathcal{R}(\widehat{\boldsymbol{\alpha}}_{2,\mathsf{real}}) = 1,$$
$$\lim_{n\to\infty} \mathcal{C}(\widehat{\boldsymbol{\alpha}}_{\mathsf{SVM}}) = 0.$$

2. *For* $(1 - r) + \frac{(p-1)}{2} < q \leq (p - r)$*, we have*

$$\lim_{n\to\infty} \mathcal{R}(\widehat{\boldsymbol{\alpha}}_{2,\mathsf{real}}) = 1,$$
$$\lim_{n\to\infty} \mathcal{C}(\widehat{\boldsymbol{\alpha}}_{\mathsf{SVM}}) = \frac{1}{2}.$$

Observe that Corollary 16 directly follows from plugging in the condition required in the bi-level ensemble for all training points usually becoming support vectors (Equation (13)), and noting that for $p > 2$, we have

$$(1 - r) + \frac{(p-1)}{2} > (1 - r) + \frac{1}{2} = \left(\frac{3}{2} - r\right).$$

Importantly, we have identified that even highly overparameterized regimes, in which all training points become support vectors, can yield good generalization for classification tasks when the hard-margin SVM is used.

5.2 Path to analysis: Classification vs regression test error

The first step to proving Theorem 13 is obtaining clean expressions for both classification and regression test error. The 1-sparsity assumption that we have made on the unknown signal enables us to do this as a function of natural quantities corresponding to the preservation of the true feature (*survival*) and the pollution due to false features (*contamination*). If we assume that the real labels are generated by the t^{th} feature, α_t^*, then we can define these quantities for any solution $\widehat{\boldsymbol{\alpha}}$. First, as classically observed in statistical signal processing, the estimated coefficient corresponding to the true feature α_t^* will experience *shrinkage* and be attenuated by a factor that we denote as *survival*. From Assumption 1, we defined $\boldsymbol{\alpha}^* := \frac{1}{\sqrt{\lambda_t}} \cdot \mathbf{e}_t$, and so we have

$$\mathsf{SU}(\widehat{\boldsymbol{\alpha}}, t) = \frac{\widehat{\alpha}_t}{\alpha_t^*} = \sqrt{\lambda_t}\widehat{\alpha}_t \tag{14}$$

Second, we have the *false discovery of features*. We measure the effect of this false discovery for prediction on a test point X by a *contamination* term:

$$B = \sum_{j=1, j\neq t}^{d} \widehat{\alpha}_j \boldsymbol{\phi}_j(\mathbf{X}). \tag{15}$$

Recall that X is random, and the features $\boldsymbol{\phi}(X)$ are zero-mean. Therefore, B is a zero-mean random variable. Accordingly, we can define the standard deviation of the contamination term on a test point as below:

$$\begin{aligned}\mathsf{CN}(\widehat{\boldsymbol{\alpha}}, t) &= \sqrt{\mathbb{E}\left[B^2\right]} \\ &= \sqrt{\sum_{j=1, j\neq t}^{d} \lambda_j \widehat{\alpha}_j^2}. \end{aligned} \tag{16}$$

where the last step follows from the orthogonality of the d features. The ideas of survival and contamination can be related to the classical signal processing concept of *aliasing*; Figure 7 in Appendix A provides an illustration.

We state and prove the following proposition, which directly expresses regression and classification test loss in terms of these terms.

Proposition 17 *Under the* 1*-sparse noiseless linear model, the regression test loss (excess MSE) is given by:*

$$\mathcal{R}(\widehat{\boldsymbol{\alpha}}) = (1 - \mathsf{SU}(\widehat{\boldsymbol{\alpha}}, t))^2 + \mathsf{CN}^2(\widehat{\boldsymbol{\alpha}}, t). \tag{17}$$

and the classification test loss (excess classification error) is given by:

$$\mathcal{C}(\widehat{\boldsymbol{\alpha}}) = \frac{1}{2} - \frac{1}{\pi}\mathsf{tan}^{-1}\left(\frac{\mathsf{SU}(\widehat{\boldsymbol{\alpha}}, t)}{\mathsf{CN}(\widehat{\boldsymbol{\alpha}}, t)}\right). \tag{18}$$

We can think of the quantity $\mathsf{SU}(\widehat{\boldsymbol{\alpha}}, t)/\mathsf{CN}(\widehat{\boldsymbol{\alpha}}, t)$ *as the effective "signal-to-noise ratio" for classification problems.*

Proof We first prove Equation (17). Recall that for any estimator $\widehat{\boldsymbol{\alpha}}$, the excess MSE is given by

$$\begin{aligned}\mathcal{R}(\widehat{\boldsymbol{\alpha}}) :&= \mathbb{E}[(\langle \boldsymbol{\phi}(X),\, \boldsymbol{\alpha}^* - \widehat{\boldsymbol{\alpha}}\rangle)^2] \\ &= \sum_{j=1}^{d} \lambda_j (\alpha_j^* - \widehat{\alpha}_j)^2,\end{aligned}$$

and then substituting in the 1-sparse Assumption 1 gives us Equation (17).

Next, we prove Equation (18). Since $\boldsymbol{\phi}(X) = \boldsymbol{\Sigma}^{1/2}\mathbf{W}$ for $\mathbf{W} = (W_1, \ldots, W_d) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, we can write $\boldsymbol{\phi}(X)^\top \boldsymbol{\alpha}^* = W_t$ and $\boldsymbol{\phi}(X)^\top \widehat{\boldsymbol{\alpha}} = \sum_{j=1}^{d} \sqrt{\lambda_j} W_j \widehat{\alpha}_j$. Thus, the excess classification error of $\widehat{\boldsymbol{\alpha}}$ is given by

$$\mathcal{C}(\widehat{\boldsymbol{\alpha}}) = \mathbb{P}\left(\boldsymbol{\phi}(X)^\top \widehat{\boldsymbol{\alpha}} \boldsymbol{\phi}(X)^\top \boldsymbol{\alpha}^* \leq 0\right) = \mathbb{P}\left(\sqrt{\lambda_t}\widehat{\alpha}_t W_t^2 + W_t \cdot \sum_{j\neq t} \sqrt{\lambda_j}\widehat{\alpha}_j W_j \leq 0\right).$$

Now, the random sum $\sum_{j\neq t} \sqrt{\lambda_j}\widehat{\alpha}_j W_j$ has a Gaussian distribution with mean zero and variance $\mathsf{CN}(\widehat{\boldsymbol{\alpha}}, t)^2$. Since the $\{W_j\}_{j=1}^d$ are independent, the classification test error of $\widehat{\boldsymbol{\alpha}}$ is the probability of the following event:

$$\mathsf{SU}(\widehat{\boldsymbol{\alpha}}, t) U^2 + U \cdot \mathsf{CN}(\widehat{\boldsymbol{\alpha}}, t) V \leq 0,$$

where U and V are independent standard Gaussian random variables. This event is equivalently written as

$$\frac{V}{U} \leq -\frac{\mathsf{SU}(\widehat{\boldsymbol{\alpha}}, t)}{\mathsf{CN}(\widehat{\boldsymbol{\alpha}}, t)}.$$

Since V/U follows the standard Cauchy distribution with cumulative distribution function $F(t) = \frac{1}{2} + \frac{1}{\pi}\mathsf{tan}^{-1}(t)$, the claim follows. ■

Equations (17) and (18) give us an initial clue as to why classification test error can be easier to minimize than regression test error. For the right hand side of Equation (17) to be small, we need $\mathsf{SU} \to 1$ to avoid shrinkage, as well as $\mathsf{CN} \to 0$ to avoid contamination. However, for the right hand side of Equation (18) to be small, we only require the ratio of contamination to survival to be small (i.e. $\mathsf{CN}/\mathsf{SU} \to 0$). Clearly, the former condition directly implies the latter, showing that classification is "easier" than regression[9]. Theorem 13 is proved fully in Appendices D and E in the following series of steps:

1. Matching (non-asymptotic) upper and lower bounds are proved on both survival and contamination for interpolation of both real and binary labels. The full statements for these bounds are contained in Theorems 22 and 23 in Appendix D.1.

2. These bounds are substituted into the bi-level ensemble to get asymptotic scalings for classification and regression test error (Appendix E).

The bulk of the technical work is involved in proving the matching bounds on survival and contamination, i.e. Theorems 22 and 23. Although these results are inspired by the calculations provided in Appendix A.2 for the Fourier case, we build on the techniques provided in Bartlett et al. (2020) for Gaussian features, particularly making use of fundamental concentration bounds that were proved on "leave-one-out" matrices in that work. We build on these techniques to sharply bound both the "survival" and "contamination" terms, and thus obtain matching upper and lower bounds for the classification test error. Crucially, our analysis needs to circumvent issues that stem from effective misspecification in the linear model that arise from the sign operator. While we do not provide a generic analysis of "misspecification noise," we exploit the special misspecification induced by the sign operator in a number of technical equivalents of the aforementioned random matrix concentration results.

We essentially show that this induced misspecification makes no difference, asymptotically, to classification error arising from interpolation from binary labels, and the behavior is essentially the same as though we had instead interpolated the real output. This is another interesting consequence of requiring only the ratio $\frac{\mathsf{CN}}{\mathsf{SU}} \to 0$, as opposed to the stronger requirements for regression, $\mathsf{CN} \to 0$ and $\mathsf{SU} \to 1$. We will see in Appendix E that in the asymptotic limit $n \to \infty$, interpolation of binary *noiseless* labels attenuates the signal by a factor exactly equal to $\sqrt{\frac{2}{\pi}}$. This also corresponds to the attenuation factor of signal that has been traditionally been observed as a result of 1-bit quantization applied before

9. Our decomposition of classification error is reminiscent of the decomposition by Friedman (1997) into the ratio of terms depending on the variance (like contamination) and bias (like survival) respectively. Because our data is Gaussian, Proposition 17 allows an *exact* decomposition.

a matched filter[10] (Turin, 1976; Chang, 1982). Since this factor is strictly positive, it does not affect the asymptotic classification error.

In fact, the non-asymptotic scalings of survival and contamination terms are unaffected even by non-zero label noise on classification training data, provided that the label noise still preserves non-trivial information about the signal. The survival is further attenuated by a non-zero factor of $(1 - 2\nu^*)$, which is strictly positive as long as $\nu^* < 1/2$. Observe that this is equivalent to a hypothetical scenario where the binary labels take on "shrunk" values $\{-(1 - 2\nu^*), (1 - 2\nu^*)\}$ instead of the usual $\{-1, 1\}$. As long as $\nu^* < 1/2$, the magnitude of the labels is strictly non-zero and so the labels still provide useful information for classification.

Finally, it is natural to ask how fundamental our assumptions of Gaussianity on data and bi-level covariance structure are to our main generalization result (Theorem 13). We chose the bi-level ensemble to illustrate the separation between classification and regression in the cleanest possible way. However, Theorems 22 and 23 do provide non-asymptotic expressions for survival and contamination for *arbitrary* covariance matrices. In principle, these expressions can be plugged into Proposition 17 to get upper and lower bounds on classification error for arbitrary covariance matrices. Further, the analysis of benign overfitting in linear regression (Bartlett et al., 2020; Muthukumar et al., 2020) extends to sub-Gaussian features. In the same spirit, we can show that the results — including the existence of the intermediate regime, in which classification works but regression does not — extend to a weaker assumption of *independence* and sub-Gaussianity on the underlying features. This extension uses an argument similar to the Fourier-case argument given in Appendix A.2 but requires a more direct treatment of the approximation error arising from misspecification. We provide this argument in a forthcoming note. Our results *do not* extend to kernel settings, where there can be complex dependencies among the (infinite-dimensional) features. This is an important direction for future work.

6. Examining margin-based explanations for generalization

In this section, we explore the potential for generalization bounds as a function of training data margin to explain the behavior we have observed for classification tasks in the overparameterized regime. Through simple experiments, we demonstrate that margin-based generalization bounds are uninformative in sufficiently overparameterized settings.

6.1 The historical role of margin

For a particular function class $\mathcal{F}$, *uniform convergence bounds* conservatively approximate the generalization error of $f \in \mathcal{F}$ by that of the least generalizable function in $\mathcal{F}$. The ensuing generalization bounds typically depend on measures of complexity, such as the Vapnik-Chervonenkis dimension, which increase with the number of parameters in the model. Thus, the uniform convergence approximation is not as good when $\mathcal{F}$ is large, e.g. the model has several parameters. This shortcoming of uniform convergence-based bounds was first brought into focus by the remarkable success of boosting with a very large

10. Recall that Muthukumar et al. (2020) naturally connected matched filtering to minimum-ℓ_2-norm interpolation.

number of primitive classifiers (Schapire et al., 1998). The main observation was that even after the training 0-1 loss became zero, increasing the number of primitive classifiers in the boosted model still reduced the test error.

An analysis in terms of the *training data margin* was proposed as a possible explanation for this behavior for classifiers $f(\cdot)$ that make their predictions by discretizing the outputs of a real-valued function $g \in \mathcal{G}$, i.e. $f(X) = \mathsf{sgn}(g(X))$. The training margin, $\gamma := \min_i Y_i g(X_i)$ can be intuitively interpreted as a measure of prediction confidence; for linear classifiers, it is precisely the minimum (over training points) distance to the decision boundary. The worst-case margin is not the only quantity that has been considered: generalization bounds based on a weighted combination of margin on all training data points have also been considered and demonstrated to be sharper in certain settings (Gao and Zhou, 2013). In the settings we investigate, all training data points become support vectors – therefore *the margins at each training point are equal*, and all such notions of margin become equivalent.

Under certain conditions, margin-based generalization bounds can scale far slower with the number of parameters in the model than uniform convergence bounds; for example, in boosting, the dependence is reduced to ln(# of primitive classifiers). Since the margin γ could be artificially increased (without changing any of the predictions) simply by rescaling the real-valued function $g(\cdot)$, the quantity of interest is an appropriately *normalized* margin, e.g. the margin normalized by the Lipschitz constant of the learned function $g(\cdot)$ or its approximation.

The decrease of generalization error despite increasing complexity in "modern" overparameterized regimes is strongly reminiscent of the observations from boosting with a large number of primitive classifiers. It is of particular interest to examine the ensuing generalization bounds for the hard-margin SVM, which maximizes margin on linearly separable data. For the case of linear classifiers, the normalized margin is defined as $\gamma_N = \frac{\gamma}{\|\widehat{\boldsymbol{\alpha}}\|_2}$. We can now state the ensuing classification test error (i.e. 0-1 test loss) as a function of the normalized margin. Notation in the statement is adapted to be consistent with the notation in this paper — for an elementary verification, see Appendix H.

Theorem 18 (Theorem 21, (Bartlett and Mendelson, 2002)) *For a random test point (X, Y) drawn from the same distribution as the training data, the following holds with probability $(1-\delta)$ over the training data $\{X_i, Y_i\}_{i=1}^n$:*

$$\Pr[\mathsf{sgn}(\widehat{\boldsymbol{\alpha}}^T \boldsymbol{\phi}(X)) \neq Y] \leq \frac{1}{n}\sum_{i=1}^{n} l_\gamma(\widehat{\boldsymbol{\alpha}}^\top \boldsymbol{\phi}(X_i)\cdot Y_i) + \frac{4}{\gamma_N}\cdot\frac{\|\boldsymbol{\Phi}_{\mathsf{train}}\|_{\mathsf{F}}}{n} + \left(\frac{8}{\gamma}+1\right)\cdot\sqrt{\frac{\ln(4/\delta)}{2n}} \quad (19)$$

where the ramp loss function l_γ is defined by

$$l_\gamma(z) := \begin{cases} 1 \text{ if } z \leq 0 \\ 1 - \frac{z}{\gamma} \text{ if } 0 < z \leq \gamma \\ 0 \text{ if } z > \gamma. \end{cases}$$

When the training data are separable, we apply the above bound setting the first (average training loss) term to 0, and only consider the the second term in the bound, i.e. we ignore

the high-probability term. Equation (19) reminds us that there is a critical dependence on the intrinsic data dimension, captured by the term $\|\mathbf{\Phi}_{\mathsf{train}}\|_{\mathsf{F}}$. We will shortly see that this dependence is critical to track in the overparameterized regime.

6.2 Can margin track performance of overparameterized models?

We now investigate whether this generalization bound is effective in tracking the true test classification error for the hard-margin SVM in our setting for a number of choices of featurization. Importantly, we consider the solution $\widehat{\boldsymbol{\alpha}}_{\mathsf{SVM}}$ only in sufficiently overparameterized settings under which all training points become support vectors with high probability; therefore, the un-normalized margin $\gamma = 1$ and the normalized margin of the SVM solution is exactly equal to $\gamma_N = \frac{1}{\|\widehat{\boldsymbol{\alpha}}\|_2}$.

We study the evolution of margin, the ensuing upper bound in Equation (19), and the true test classification error as we increase the level of overparameterization for two choices of featurizations: isotropic Gaussian features (Definition 4) which generalize poorly according to Theorem 13 and weak features (Definition 9), which are known to exhibit the double-descent behavior. For the case of isotropic features, we retain our 1-sparse assumption from Section 5. For the case of weak features, we consider $Y_i = \mathsf{sgn}(U_i)$ for $i \in \{1, \ldots, n\}$.

Figure 5 plots the isotropic case, and Figure 6 plots the weak features case. For both figures, we hold the number of training points, n constant and vary the number of features, d, to tune the extent of overparameterization. In both Figures 5(a) and 6(a), the normalized margin increases with increasing d, since the optimizer can use more features to meet the constraint in Equation (3a). The generalization bounds in Figure 5(b) and Figure 6(b) are consequently very similar as well. However, while the test classification loss increases with d for isotropic features, it decreases with d for weak features.

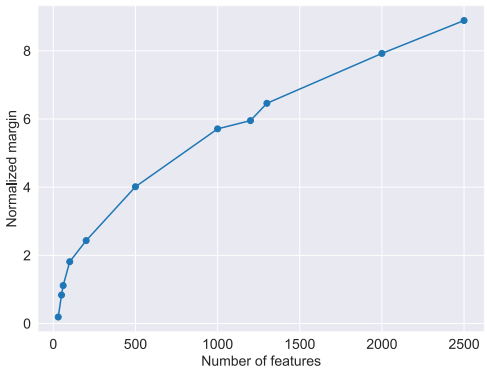
Figures 5 and 6 together show that the relationship between margin and generalization is more complex than typically assumed in highly overparameterized regimes. We highlight a few observations:

1. In both featurizations, the generalization bounds are always greater than 1, and hence, tautological. A similar empirical observation was made by Dziugaite and Roy (2017) for a two-layer neural network. The $\|\mathbf{\Phi}_{\mathsf{train}}\|_{\mathsf{F}}$ term, which represents the scale of the data, effectively cancels out any beneficial effect of increasing normalized margin[11]. Intuitively, it is clear that *feature-space* margin-based bounds will have to scale with the intrinsic input dimension, which itself is overparameterized for Gaussian featurization.

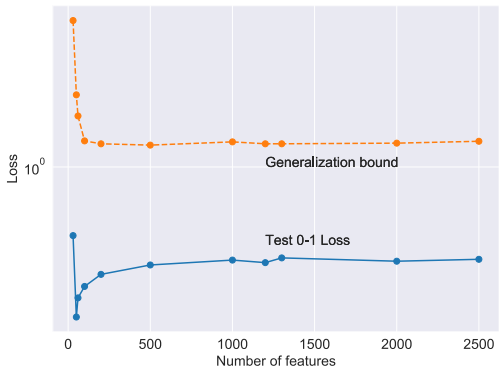
2. Whether margin is *qualitatively* predictive of generalization is also unclear, as evidenced by the contrasting examples of weak features and isotropy. Under both featurizations, the normalized margin increases with increased overparameterization; but the actual test error behaves very differently (decreasing for weak features, but increasing for isotropy).

Thus, we see that margin-based bounds are not predictive of the behavior of overparameterized models in our setting. It is still possible that an appropriate sense of large margin

11. In fact, for the case of minimum-ℓ_2-norm interpolation and isotropic features, this can be verified quantitatively, as we know that $||\widehat{\boldsymbol{\alpha}}||_2 \sim \sqrt{\frac{n}{d}}$ and $||\mathbf{\Phi}_{\mathsf{train}}||_{\mathsf{F}} \sim \sqrt{nd}$ with high probability.

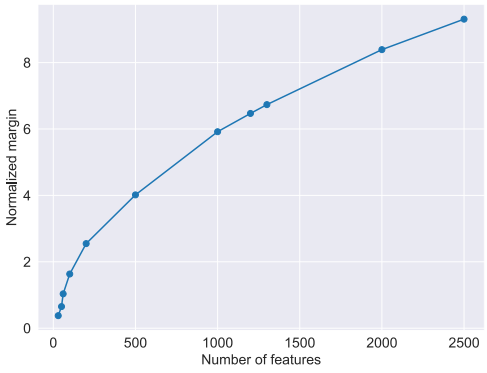


(a) Normalized margin.

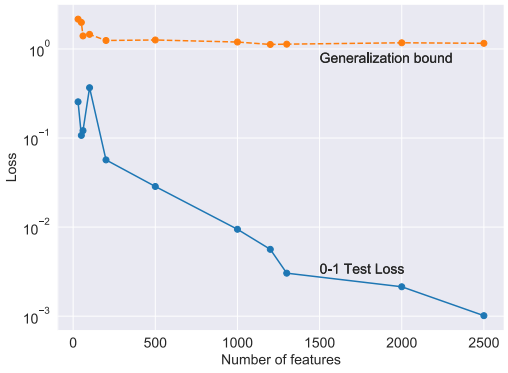


(b) Comparison of the generalization bound (19) with true test classification loss.

Figure 5: Evolution of normalized margin, ensuing generalization bound and true classification test loss as a function of number of features d for isotropic Gaussian features ($n = 32$ fixed). Observe that the terms $\|\mathbf{\Phi}_{\mathsf{train}}\|_{\mathsf{F}}$ and $\|\widehat{\boldsymbol{\alpha}}\|_2$ cancel each other's effect on the bound, leading to a roughly constant bound. The true test error increases as d is increased.



(a) Normalized margin.



(b) Comparison of the generalization bound (19) with true test classification loss.

Figure 6: Evolution of normalized margin, ensuing generalization bound and true classification test loss as a function of number of features d for weak features in Definition 9 ($n = 32$ fixed, $\sigma = 0.1$). Observe that the terms $\|\mathbf{\Phi}_{\mathsf{train}}\|_{\mathsf{F}}$ and $\|\widehat{\boldsymbol{\alpha}}\|_2$ cancel each other's effect on the bound, leading to a roughly constant bound. The true test error decreases as d is increased.

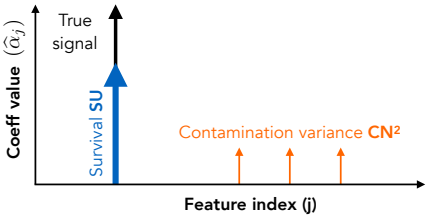
implies good generalization in certain cases. In particular, for linear models, maximizing the margin is equivalent to minimizing the norm — which, as we have seen, has important generalization properties. However, evidence of this needs to come from first-principles analysis, not from the existing bounds.

The recent work of Negrea et al. (2019) suggests a way forward in the interpolating regime via the introduction of appropriate surrogates, that implicitly capture the good generalization properties vis-a-vis the underlying patterns and the learning algorithm used. It would be interesting to see how these ideas could be unified with the survival/contamination perspective developed here across all three regimes identified in Theorem 13.

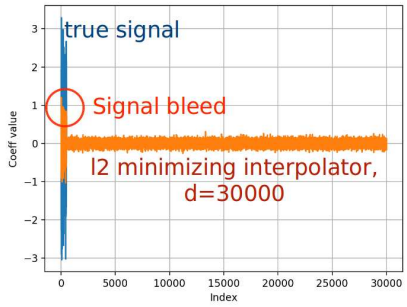
Acknowledgments

We would like to acknowledge the Simons Institute Summer 2019 program on "Foundations of Deep Learning", which facilitated the initial collaboration that led to this work. More broadly, we thank the participants of this program for many stimulating research discussions that inspired this collaboration — especially Suriya Gunasekar and Matus Telgarsky.

MB acknowledges federal support from NSF and a Google Research Award. DH acknowledges support from NSF grant CCF-1740833 and a Sloan Research Fellowship. AS acknowledges the support of the ML4Wireless center member companies and NSF grants AST-144078 and ECCS-1343398.



(a) Illustration for 1-sparse signal (Assumption 1).



(b) Actual signal "bleed" and "contamination" for 30000 isotropic Gaussian features, and 1000 samples of 500-sparse signal.

Figure 7: Illustrations of survival and contamination factors that affect both classification and regression test error. Here, signal "bleed" refers to the qualitative phenomenon where magnitude of recovered components of the true signal is much smaller than the magnitude of true signal components and is an extension of the survival concept to the hard-sparse setting.

Appendix A. Fourier features on regularly spaced training data: An "ultra-toy" model

In Muthukumar et al. (2020), the case of Fourier features on regularly spaced training data was introduced and studied as an "ultra-toy", or caricature model to highlight the consequences of overparameterization in linear regression on noisy data. The ramifications of ℓ_2-minimization are clearly illustrated through this model, as an explicit connection can be made to the classical phenomenon of *aliasing* that is involved to understand the undersampling of continuous time signals. Using this signal processing perspective, survival and contamination are natural quantities of interest, as illustrated in Figure 7(a) for the 1-sparse case. In Figure 7(b), we see how these concepts would *qualitatively* manifest more generally when the underlying signal is hard-sparse.

As we illustrate in this section, appropriate weightings of these features under this "ultra-toy" model also helped us conjecture all of the main results of this paper.

The Fourier ensemble is defined below.

Definition 19 (Weighted Fourier features on regularly spaced data) *We consider n (odd) regularly spaced training points from $(-\pi, +\pi)$ — specifically the sequence $(-\pi + \frac{\pi}{n}, -\pi + \frac{3\pi}{n}, \ldots, -\frac{2\pi}{n}, 0, +\frac{2\pi}{n}, \ldots, +\pi - \frac{\pi}{n})$, a test distribution of X drawn uniformly at random from $(-\pi, +\pi)$, and the d (odd multiple of n) features chosen to be the standard real orthonormal Fourier features:*

$$\frac{1}{\sqrt{2\pi}}, \frac{1}{\sqrt{\pi}}\sin(x), \frac{1}{\sqrt{\pi}}\cos(x), \ldots, \frac{1}{\sqrt{\pi}}\sin\left(\frac{d-1}{2}x\right), \frac{1}{\sqrt{\pi}}\cos\left(\frac{d-1}{2}x\right).$$

For doing interpolative inference using a weighted norm minimization, we define the weights[12] *corresponding to sines and cosines of frequency j by $\{\lambda_j\}_{j=0}^{\frac{(d-1)}{2}}$. Following the convention of the rest of the paper, we take the weights $\{\lambda_j\}$ to be a decreasing, strictly positive sequence.*

Exact aliases are defined as distinct features that agree with each other (possibly up to a constant multiple) on all the sampled points. The Fourier featurization allows exact aliases to exist. There are three different groups of these exact aliases:

- The initial constant feature is essentially aliased by the cosines at every multiple[13] of n.
- Each cosine feature in the first n features (namely corresponding to a frequency $j \in \{1, 2, \ldots, \frac{n-1}{2}\}$) picks up $(\frac{d}{n} - 1)$ cosine aliases with frequencies $(n-j), (n+j), (2n-j), (2n+j), \ldots$. This is because cosine is an even function and the training samples are symmetrically distributed about 0.
- Similarly, each sine feature in the first n features (corresponding to a frequency $j \in \{1, 2, \ldots, \frac{n-1}{2}\}$) picks up $(\frac{d}{n} - 1)$ sine aliases with frequencies $(n-j), (n+j), (2n-j), (2n+j), \ldots$. However, because sine is an odd function, these aliases have their signs alternating with the $(kn-j)$ ones being multiplied by (-1) and the $(kn+j)$ ones being exact aliases.

A.1 Empirical evidence that all training points become support vectors

Consider, first, the case of $d = n$ and $\lambda_j = 1$ for all j. The orthonormality of the Fourier features above essentially (this argument would be exact if we used the complex Fourier features) makes the first n columns of the training data feature matrix look like a rotation of a scaled version of the identity. In such a situation, minimizing the ℓ_2-norm of the learned parameters in a hard-margin SVM essentially forces[14] every point to be a support vector. Adding more aliases (in a balanced way) for all features is not going to change this. This leads to every point becoming a support vector in the isotropic case for any d that is a multiple of n.

The case of non-isotropic overparameterized models is more complex. Here, we describe the experiment underlying Figure 1 in the main text that first showed that all training points became support vectors in high dimensions (Theorem 11). We conducted this experiment for regularly spaced training data and the Fourier featurization as defined above, with

12. If α_j represents the learned coefficient on the cosine at frequency f, and β_j the learned coefficient on the sine at frequency f, the minimization is of $\sum_j \frac{\alpha_j^2+\beta_j^2}{\lambda_j}$. A higher λ_j means that frequency is favored.
13. For ease of exposition, the minor issue of the constant feature having a slightly different scaling vis-a-vis its aliases is going to be ignored in this treatment, but this is simply a matter of keeping track of notation. Alternatively, we could eliminate this by using complex Fourier features. We will finesse this issue here by simply not allowing the true signal to have a constant term in it.
14. Why? Suppose the scalar prediction on one of the training points was larger than +1. If so, we could reduce the norm of the learned parameters without impacting any other training point's prediction by making this point have a scalar prediction of +1. Norm-minimization and a full complement of orthonormal vectors in the training matrix forces every point to be a support vector. This can alternatively be viewed as a consequence of extreme symmetry — under a Fourier featurization and the 2-norm, no training point is special if they are regularly spaced.

polynomial decay in the weights used, i.e. $\lambda_k = \frac{1}{k^m}$ for $m \geq 0$, and (since the training data $\{X_i\}_{i=1}^n$ is 1-dimensional), visualized both the minimum-ℓ_2-norm interpolation and the SVM. Figure 1 in the main text shows a remarkable equivalence between the two solutions for various values of m, and provides initial empirical evidence for this phenomenon (that we theoretically established for Gaussian featurization in Theorem 11).

A.2 Regression vs Classification

To see the counterpart of Theorem 13, which compares classification and regression test error of interpolating solutions, we consider the underlying true function to be $\cos(x)$. At training time, we get actual real-valued outputs $z_j = \cos(x_j)$ corresponding to the n regularly spaced points $\{x_j\}$. The minimum-ℓ_2-norm interpolation of real-valued output leads to the following coefficients on the d Fourier features:

$$\widehat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha} \,\big|\, \boldsymbol{\Phi}_{\mathsf{train}}\boldsymbol{\alpha} = \mathbf{z}_{\mathrm{train}}}{\arg\min} \left(\frac{1}{\lambda_H} \sum_{j=0}^{s-1} \alpha_j^2 \right) + \sum_{j=s}^{d-1} \alpha_j^2 \tag{20}$$

Because of the presence of *exact aliases*, the training data matrix $\boldsymbol{\Phi}_{\mathsf{train}}$ consists of n distinct columns that repeat again and again. In keeping with the bi-level covariance model in Definition 5, we scale the parameters (s, λ_H, d) with n in a coordinated way. Recall that the number of prioritized features is given by $s := n^r$ for $r \in [0, 1)$, and the number of features $d = n + n^p$ for $p > 1$. (We added an extra term of n to make it easier to count the aliases. This has no asymptotic effect when $p > 1$ and $n \to \infty$.) The λ_H represents how much we favor the special features and in keeping with the scaling in Definition 5, we set $\lambda_H = n^{p-r-q}$ for some $q \in [0, p - r]$.

Because of the known orthogonality of the sine and cosine features on n regularly spaced points, the first n columns of $\boldsymbol{\Phi}_{\mathsf{train}}$ are orthogonal. This means that the solution $\widehat{\boldsymbol{\alpha}}$ will only have nonzero entries in the positions that correspond to the $\frac{d}{n} = 1 + n^{p-1}$ different columns of $\boldsymbol{\Phi}_{\mathsf{train}}$ that are copies of the column corresponding to the feature $\cos(x)$. Since $s < n$, exactly one of these will be favored and so the optimization problem in Equation (20) turns into the much simpler problem:

$$\min_{a,b \,\big|\, a + n^{p-1}b = 1} \frac{a^2}{n^{p-r-q}} + n^{p-1}b^2 \tag{21}$$

where a represents the recovered coefficient corresponding to the true underlying feature $\cos(x)$ and b represents the coefficients on all of its exact aliases.

An elementary calculus calculation[15] shows that Equation (21) is solved by:

$$a = \frac{\lambda_H}{\lambda_H + \left(\frac{d}{n} - 1\right)} = \frac{1}{1 + n^{q-(1-r)}} \text{ and} \tag{22a}$$

$$b = \frac{1}{\lambda_H + \left(\frac{d}{n} - 1\right)} = \frac{1}{n^{p-r-q} + n^{p-1}}. \tag{22b}$$

15. See the appendix of Muthukumar et al. (2020) for more discussion of this calculation and its connection to matched filtering in signal processing.

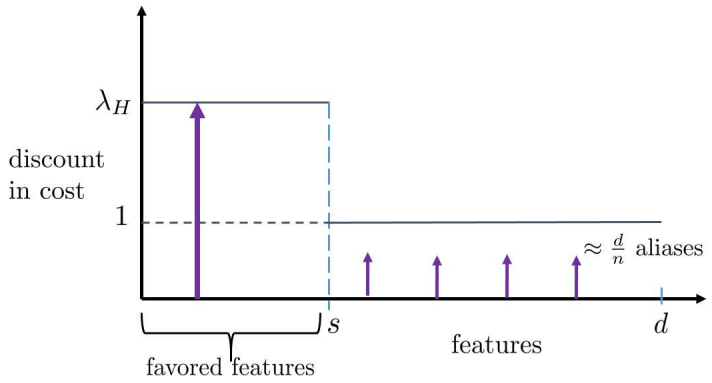


Figure 8: An illustration of the bi-level model for the Fourier features.

The reader can verify that a represents the survival of the true signal. For large enough n, this is approximated[16] by

$$a \approx \begin{cases} 1 & \text{if } q < 1-r \\ n^{-(q-(1-r))} & \text{if } q > 1-r \end{cases}. \tag{23}$$

This approximation is guaranteed to be good to within a factor of 2 everywhere and is usually much better. Notice that Equation (23) is the Fourier-feature counterpart of the upper and lower bounds on survival in Lemmas 32 (binary labels) and 33 (real-valued output). Now, taking $n \to \infty$, we get

$$a_\infty = \begin{cases} 1 & \text{if } q < (1-r) \\ 0 & \text{if } q > (1-r) \end{cases} \tag{24}$$

which shows that the signal only fully survives if $q < (1-r)$.

Let us now measure the *contaminating* effect of falsely discovered features. Following Equation (15), we denote $B(\mathbf{X})$ as the random variable that represents the contribution of all of the aliases to the predictions. Each of the Fourier features of non-zero frequency is zero-mean and has variance 1. From the orthonormality (in expectation over test data) of the aliases, we get

$$\begin{aligned} \mathrm{Var}[B(\mathbf{X})] &= n^{p-1} b^2 \\ &= \left(\frac{1}{n^{\frac{p}{2}+\frac{1}{2}-r-q} + n^{\frac{(p-1)}{2}}} \right)^2, \end{aligned} \tag{25}$$

where in the last step, we substituted Equation (22b). Notice that $\frac{(p-1)}{2} > \frac{p}{2} + \frac{1}{2} - r - q$ whenever $q > (1-r)$, and so asymptotically we get

$$\mathsf{CN} = \sigma_{CN} \approx \begin{cases} n^{-\left(\frac{p+1}{2} - (q+r)\right)} & \text{if } q < (1-r) \\ n^{-\frac{(p-1)}{2}} & \text{if } q > (1-r) \end{cases} \tag{26}$$

16. In the style of the Bode Plot of a one-pole low pass filter.

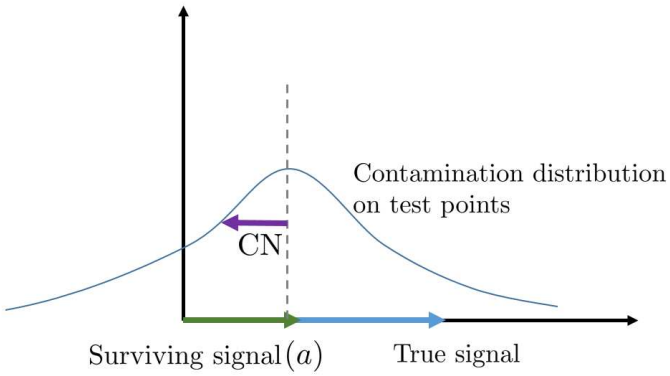


Figure 9: Illustration of how contamination can flip the sign of the prediction at a test point. The survival *relative* to the standard deviation of the contamination, CN, is what matters — if the latter is much smaller than the former, then the probability of classification error is low.

This approximation is guaranteed to be good to within a factor of 2 everywhere and is usually much better. Notice that this expression is the Fourier-feature counterpart of the lower bound on contamination established for Gaussian features in Lemma 36.

Thus, provided that $q < (1-r)$, the expression in Equation (26) always decays to zero as $n \to \infty$, regardless of which case we are in. The combination of Equations (26) and (24) tells us that regression in this problem can work[17] to get mean-square-error approaching zero as long as $q < (1-r)$. On the other hand, when $q > (1-r)$, signal does not asymptotically survive and regression MSE approaches the null risk.

A.2.1 Implications for classification: existence of the separating regime

For classification, we only care about predicting $\mathsf{sgn}(\cos(\mathbf{X}))$ correctly with high probability when $\mathbf{X} \sim \text{Unif}[-\pi, \pi]$. Clearly, classification also works under the conditions for which regression works (i.e. $q < (1-r)$), but, as we showed in Theorem 13, can work even in the absence of these conditions. Recall that when $q > (1-r)$, the survival factor $a \to 0$ as $n \to \infty$. However, if the contamination is small enough, i.e. $\sigma_{CN} \ll a$, the

17. In fact, the argument also works for *noisy training* data, i.e. $\mathbf{Z}_i = \sin(x_i) + \mathbf{W}_i$ where the noise $\mathbf{W}_i$ is iid and has zero mean, bounded support, and finite variance σ^2. The formal argument is in Muthukumar et al. (2020), but is summarized here for the ease of the reader. From the central limit theorem and the theory of wide-sense-stationary random variables, in the limit, the representation of the noise part will look marginally Gaussian in the basis of the first n columns of $\mathbf{\Phi}_{\mathsf{train}}$, where each of them will be $\mathcal{N}\left(0, \frac{\sigma^2}{n}\right)$. The first s of these will survive and thereby contribute a variance of approximately $\sigma^2 n^{r-1} a^2$ to test points, while the other $(n-s)$ of these will be absorbed across all the aliases and thereby contribute a contamination variance of $\sigma^2 n^{-p}$. The total contribution will be dominated by the $\sigma^2 n^{-(1-r)} a^2$ term. If $q < (1-r)$ and $a \approx 1$, this term vanishes as $n \to \infty$ and so additive noise does not contribute to regression error unless the noise variance σ^2 itself grows with n (at a rate faster than n^{1-r}).

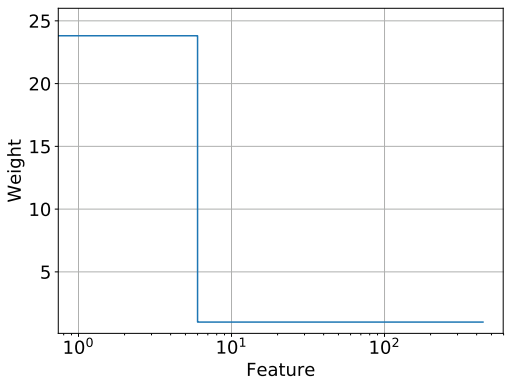
probability of classification error is extremely low, as illustrated in Figure 9. We observe from Equations (26) and (23) that $\sigma_{CN} \ll a$ if $q < (1-r) + \frac{(p-1)}{2}$. When that happens, classification will asymptotically work.

To see this more formally, we can upper bound the expression of classification error and show that it goes to zero as $n \to \infty$ under these conditions[18]. We use a union bound together with Chebyshev's inequality in a manner reminiscent of typicality proofs in information theory (Cover and Thomas, 2012). Let $\epsilon = \frac{(p-1)}{2} - (q-(1-r))$ be the difference between the relevant two exponents of n. Then, we define the events $\mathcal{A} := \{x| \ |\cos(x)| < 2n^{-\frac{\epsilon}{2}}\}$ and $\mathcal{B} := \{x| \ |B(x)| > n^{-\frac{\epsilon}{2}} n^{-(q-(1-r))}\}$. The event $\mathcal{A}$ corresponds to having an atypically weak signal in the true feature, and the event $\mathcal{B}$ corresponds to having an atypically large contamination term. Observe that if neither event $\mathcal{A}$ nor event $\mathcal{B}$ holds, we can substitute Equation (23) to get $|a\cos(x)| \geq 2|B(x)|$, and this implies that a classification error will not occur. Therefore, the probability of classification error is *upper bounded* by $\Pr[\mathcal{A} \cup \mathcal{B}]$, and by the union bound it suffices to upper bound the probability of each of these events individually. We start with the "weak signal" event $\mathcal{A}$. Because $\cos(x)$ is a function that is always differentiable in the neighborhood where $\cos(x) = 0$, this means that $\cos(\mathbf{X})$ as a random variable has a density[19] in the neighborhood of 0. Consequently, we have $\Pr[\mathcal{A}] = \int_{-n^{-\frac{\epsilon}{2}}}^{+n^{-\frac{\epsilon}{2}}} \frac{1}{\pi\sqrt{1-y^2}} dy = \frac{2}{\pi}\sin^{-1}(n^{-\frac{\epsilon}{2}})$ which goes to zero as $n \to \infty$. We now turn to the "unusually large contamination" event $\mathcal{B}$. Because $q < (1-r) + \frac{(p-1)}{2}$, we have $\Pr[\mathcal{B}] = \Pr\left[|B(\mathbf{X})| > n^{-\frac{\epsilon}{2}} n^{-(q-(1-r))}\right] \leq \Pr\left[|B(\mathbf{X})| > n^{\frac{\epsilon}{2}}\sigma_{CN}\right]$. By Chebyshev's inequality, we have $\Pr[\mathcal{B}] \leq n^{-\frac{\epsilon}{2}}$, which goes to zero as $n \to \infty$.

Since the probabilities of both events $\mathcal{A}$ and $\mathcal{B}$ have been shown to go to 0 as $n \to \infty$, the limiting classification error will also be zero when $q < (1-r) + \frac{(p-1)}{2}$. In fact, this argument can be extended to the case of interpolation of *binary labels* by using the Fourier series representation of the underlying true label function. Since there is now misspecification induced by the sign operator, this requires understanding the approximation-theoretic properties of the Fourier series by its first s terms as $s \to \infty$. Such a treatment is beyond the scope of this paper, and we defer it to a separate forthcoming note.

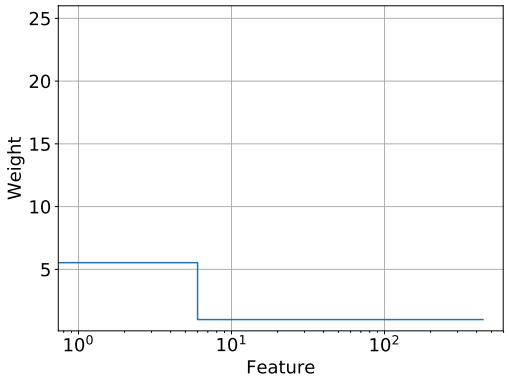
Finally, it is worth noting that the above calculation only *upper bounds* the classification error; whether this upper bound is matched by a lower bound remains an open question. This would shed light on whether, in fact, classification generalizes poorly when $q > (1-r) + \frac{(p-1)}{2}$ for the case of Fourier featurization. Figure 10 illustrates the three regimes for Fourier featurization. While the first two regimes display behavior that parallels the Gaussian-feature results in Theorem 13, the third regime is inconclusive with respect to classification performance. This is an important question for future work.

18. The *exact* Gaussian-feature expression for classification error in Proposition 17 depends solely on the ratio a/σ_{CN}. Characterizing the exact expression for Fourier features is more challenging because the contamination does not have a clean distribution, but we can upper bound the probability of classification error using the standard deviation alone.
19. This is known as a shifted arc-sine distribution.



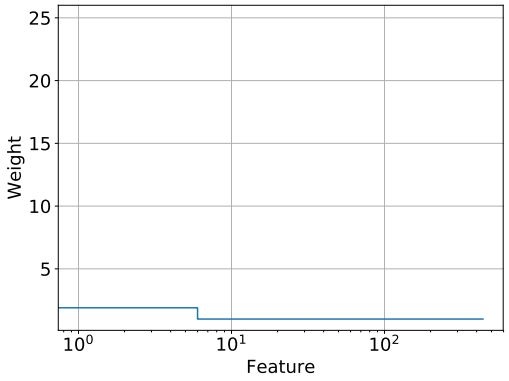
(a) $\lambda_H = 23.81$

(b) $\lambda_H = 23.81$

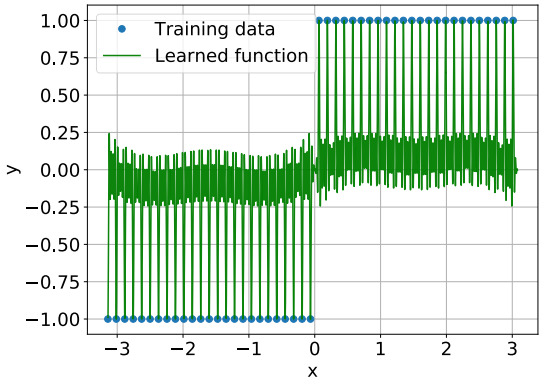


(c) $\lambda_H = 5.53$

(d) $\lambda_H = 5.53$



(e) $\lambda_H = 1.89$



(f) $\lambda_H = 1.89$

Figure 10: The three qualitative regimes illustrated using Fourier features and regularly spaced training points. The top corresponds to both regression and classification succeeding, the middle one is the intermediate regime where only classification works, and the bottom one is where neither works. Here $n = 49, s = 7, d = 441$.

Appendix B. Additional notation for proofs

We consider zero-mean Gaussian featurization, i.e. $\phi(X_i) = \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. For ease of exposition, we consider $\mathbf{\Sigma}$ to be diagonal[20]. Corresponding to a given index $t \in \{1, \dots, d\}$, we define the "leave-one-out" matrix $\mathbf{\Sigma}_{-t}$ whose eigenvalues are given by: $\mu_j(\mathbf{\Sigma}_{-t}) = \widetilde{\lambda}_j$ for $j \in \{1, \dots, d-1\}$. The relation between the spectrum $\{\widetilde{\lambda}_j\}_{j=1}^{d-1}$ and $\{\lambda_j\}_{j=1}^{d}$ is given by

$$\widetilde{\lambda}_j = \begin{cases} \lambda_j, & j < t \\ \lambda_{j+1}, & j \geq t \end{cases}. \tag{27}$$

Consider $\{\mathbf{z}_i\}_{i=1}^{d}$ i.i.d. with $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I_n})$. Observe that we can write effective Gram matrices corresponding to the full as well as the "leave-one-out" spectrum of the covariance matrix:

$$\mathbf{A} = \sum_{j=1}^{d} \lambda_j \mathbf{z}_j \mathbf{z}_j^\top = \mathbf{\Phi}_{\mathsf{train}} \mathbf{\Phi}_{\mathsf{train}}^\top, \qquad \mathbf{A}_{-t} = \sum_{j=1, j\neq t}^{d} \lambda_j \mathbf{z}_j \mathbf{z}_j^\top.$$

Using Equation (27), we can also express the "leave-one-out" Gram matrix $\mathbf{A}_{-t}$ as follows:

$$\mathbf{A}_{-t} = \sum_{j=1}^{d-1} \widetilde{\lambda}_j \mathbf{z}_j \mathbf{z}_j^\top. \tag{28}$$

We will use both of the above expressions for the leave-one-out matrix $\mathbf{A}_{-t}$ in our analysis.

Appendix C. Support vector proofs and calculations

In this section, we collect proofs and calculations that accompany Section 4, which links the SVM to the minimum-ℓ_2-norm interpolation of binary labels. We first prove Theorem 11, and then collect auxiliary calculations for the bi-level ensemble.

C.1 Proof of Theorem 11

Recall that we defined the random Gram matrix $\mathbf{A}$ as

$$\mathbf{A} := \sum_{j=1}^{d} \lambda_j \mathbf{z}_j \mathbf{z}_j^\top,$$

where $\mathbf{z}_j$ i.i.d. $\sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ reflects all the randomness in the matrix $\mathbf{A}$. Note that the spectrum $\{\lambda_j\}_{j=1}^{d}$, and all functionals of it, are deterministic.

The dual to the optimization problem (4) can be expressed as below (Boser et al., 1992):

$$\max_{\boldsymbol{\beta}} \mathbf{Y}_{\mathsf{train}}^\top \boldsymbol{\beta} - \frac{1}{2} \boldsymbol{\beta}^\top \mathbf{A} \boldsymbol{\beta}$$
$$\text{subject to}$$
$$Y_i \beta_i \geq 0 \text{ for all } i \in [n].$$

20. This is without loss of generality: if $\mathbf{\Sigma}$ were not diagonal, we could first do a coordinate transformation to the basis of the eigenvectors of $\mathbf{\Sigma}$.

Note that the *unconstrained* solution of the above is: $\boldsymbol{\beta}^* := \mathbf{A}^{-1}\mathbf{Y}_{\mathsf{train}}$. By complementary slackness, all of the constraints in the optimization problem (4) will be satisfied with equality at optimum i.e. all training points are support vectors, *if* we have

$$Y_i\beta_i^* > 0 \quad \text{for all } i \in [n]. \tag{29}$$

Thus, it suffices to establish conditions under which Equation (29) holds with high probability.

We start by showing that this is the case, provided that the condition in Equation (11) holds. To do this, we use the following lemma.

Lemma 20 *Let* $\mathbf{E} := \mathbf{A} - ||\boldsymbol{\lambda}||_1\mathbf{I}_n$. *Then, for any choice of positive constant* $0 < \epsilon < 1$ *and* $\tau > 0$, *we have (for large enough* n*),*

$$||\mathbf{E}||_{\mathsf{op}} \leq \max\{f_1(\boldsymbol{\lambda};\epsilon,\tau), f_2(\boldsymbol{\lambda};\epsilon,\tau)\} \text{ where}$$

$$f_1(\boldsymbol{\lambda};\epsilon,\tau) := \left(\frac{1}{(1-\epsilon)^2} - 1\right)||\boldsymbol{\lambda}||_1 + \frac{1}{(1-\epsilon)^2}\left(\sqrt{2||\boldsymbol{\lambda}||_2^2\left(\tau + n\ln\left(1+\frac{2}{\epsilon}\right)\right)} + 2||\boldsymbol{\lambda}||_\infty \cdot \left(\tau + n\ln(1+\frac{2}{\epsilon})\right)\right)$$

$$f_2(\boldsymbol{\lambda};\epsilon,\tau) := \left(\frac{2\epsilon}{1-\epsilon}\right)||\boldsymbol{\lambda}||_1 + \frac{1+\epsilon}{1-\epsilon}\left(\sqrt{2||\boldsymbol{\lambda}||_2^2\left(\tau + n\ln\left(1+\frac{2}{\epsilon}\right)\right)}\right) + \frac{2\epsilon}{1-\epsilon}\left(2||\boldsymbol{\lambda}||_\infty \cdot \left(\tau + n\ln\left(1+\frac{2}{\epsilon}\right)\right)\right)$$

with probability at least $(1 - 2e^{-\tau})$ *over the randomness in the matrix* $\mathbf{A}$.

Lemma 20, which essentially controls the operator norm of the error matrix $\mathbf{E}$ using a union bound with discretization (also known as the "epsilon-net" argument), is proved in Appendix F.1. Now, substituting $\tau := \ln n$ and $\epsilon := \frac{1}{36\sqrt{n}}$, all of the following inequalities can be verified to hold for large enough n:

$$\begin{aligned}
\frac{1}{(1-\epsilon)^2} - 1 &\leq \frac{1}{12\sqrt{n}} \\
\frac{2\epsilon}{1-\epsilon} &\leq \frac{2}{35\sqrt{n}} \\
\frac{1}{(1-\epsilon)^2} \cdot \sqrt{2\left(\tau + n\ln\left(1+\frac{2}{\epsilon}\right)\right)} &\leq 4\sqrt{n\ln n} \\
\frac{1+\epsilon}{1-\epsilon} \cdot \sqrt{2\left(\tau + n\ln\left(1+\frac{2}{\epsilon}\right)\right)} &\leq 4\sqrt{n\ln n} \\
\frac{2}{(1-\epsilon)^2} \cdot \left(\tau + n\ln\left(1+\frac{2}{\epsilon}\right)\right) &\leq 8n\ln n \\
\frac{4\epsilon}{1-\epsilon} \cdot \left(\tau + n\ln\left(1+\frac{2}{\epsilon}\right)\right) &\leq \frac{2n\ln n}{3\sqrt{n}}.
\end{aligned}$$

Together, this gives us both of

$$||\mathbf{E}||_{\mathsf{op}} \leq \frac{1}{12\sqrt{n}} \cdot ||\boldsymbol{\lambda}||_1 + 4\sqrt{n \ln n} \cdot ||\boldsymbol{\lambda}||_2 + 8n \ln n \cdot ||\boldsymbol{\lambda}||_\infty \text{ and}$$
$$||\mathbf{E}||_{\mathsf{op}} \leq \frac{2}{35\sqrt{n}} \cdot ||\boldsymbol{\lambda}||_1 + 4\sqrt{n \ln n} \cdot ||\boldsymbol{\lambda}||_2 + \frac{2n \ln n}{3\sqrt{n}} \cdot ||\boldsymbol{\lambda}||_\infty$$

with probability at least $(1 - 2e^{-\ln n}) = \left(1 - \frac{2}{n}\right)$. Now, observe that as a consequence of Equation (11), we have

$$||\boldsymbol{\lambda}||_2 \leq \frac{||\boldsymbol{\lambda}||_1}{72n\sqrt{\ln n}} \text{ and}$$
$$||\boldsymbol{\lambda}||_\infty \leq \frac{||\boldsymbol{\lambda}||_1}{72n\sqrt{n}\ln n}.$$

Substituting these inequalities above finally gives us

$$||\mathbf{E}||_{\mathsf{op}} \leq \left(\frac{1}{12\sqrt{n}} + \frac{4}{72\sqrt{n}} + \frac{8}{72\sqrt{n}}\right) ||\boldsymbol{\lambda}||_1 = \frac{1}{4\sqrt{n}}||\boldsymbol{\lambda}||_1 \text{ and}$$
$$||\mathbf{E}||_{\mathsf{op}} \leq \left(\frac{2}{35\sqrt{n}} + \frac{4}{72\sqrt{n}} + \frac{2}{3 \cdot 72n}\right) ||\boldsymbol{\lambda}||_1 < \frac{1}{4\sqrt{n}}||\boldsymbol{\lambda}||_1.$$

Thus, we have shown that for large enough n, we have

$$||\mathbf{E}||_{\mathsf{op}} \leq \frac{||\boldsymbol{\lambda}||_1}{4\sqrt{n}} \tag{30}$$

with probability at least $\left(1 - \frac{2}{n}\right)$.
Now, we denote $\mathbf{E}' := \frac{1}{||\boldsymbol{\lambda}||_1}\mathbf{I}_n - \mathbf{A}^{-1}$. Observe that when Equation (30) holds, we have

$$\begin{aligned}
\mu_{\mathsf{min}}(\mathbf{A}) &= \mu_{\mathsf{min}}(\mathbf{E}) + ||\boldsymbol{\lambda}||_1 \\
&\geq -||\mathbf{E}||_{\mathsf{op}} + ||\boldsymbol{\lambda}||_1 \\
&\geq -\frac{||\boldsymbol{\lambda}||_1}{4\sqrt{n}} + ||\boldsymbol{\lambda}||_1 \\
&\geq 0.9||\boldsymbol{\lambda}||_1,
\end{aligned}$$

where the last inequality again holds for large enough n. Thus, for large enough n, we have

$$\mu_{\mathsf{min}}(\mathbf{A}) \geq 0.9||\boldsymbol{\lambda}||_1. \tag{31}$$

Furthermore, since we can write $\mathbf{E}' = \frac{1}{||\boldsymbol{\lambda}||_1} \cdot \mathbf{A}^{-1} \cdot \mathbf{E}$, we get

$$\begin{aligned}
||\mathbf{E}'||_{\mathsf{op}} &\overset{\text{(i)}}{\leq} \frac{1}{||\boldsymbol{\lambda}||_1} \cdot ||\mathbf{A}^{-1}||_{\mathsf{op}} \cdot ||\mathbf{E}||_{\mathsf{op}} \\
&= \frac{1}{||\boldsymbol{\lambda}||_1 \cdot \mu_{\mathsf{min}}(\mathbf{A})} \cdot ||\mathbf{E}||_{\mathsf{op}} \\
&\overset{\text{(ii)}}{\leq} \frac{1}{0.9 \cdot ||\boldsymbol{\lambda}||_1^2} \cdot ||\mathbf{E}||_{\mathsf{op}} \\
&\overset{\text{(iii)}}{\leq} \frac{1}{0.9 \cdot ||\boldsymbol{\lambda}||_1} \cdot \frac{1}{4\sqrt{n}} \\
&\leq \frac{1}{||\boldsymbol{\lambda}||_1} \cdot \frac{1}{2\sqrt{n}},
\end{aligned}$$

where inequality (i) uses the standard inequality on product of operator norms, inequality (ii) substitutes Equation (31), and inequality (iii) substitutes Equation (30). Thus, we get, for every $i \in [n]$,

$$\begin{aligned}
Y_i \beta_i^* &= Y_i \mathbf{e}_i^\top \mathbf{A}^{-1} \mathbf{Y}_{\mathsf{train}} \\
&= Y_i \mathbf{e}_i^\top \left(\frac{1}{||\boldsymbol{\lambda}||_1} \mathbf{I}_n + \mathbf{E}' \right) \mathbf{Y}_{\mathsf{train}} \\
&= \frac{1}{||\boldsymbol{\lambda}||_1} + Y_i \mathbf{e}_i^\top \mathbf{E}' \mathbf{Y}_{\mathsf{train}} \\
&\overset{\text{(i)}}{\geq} \frac{1}{||\boldsymbol{\lambda}||_1} - \sqrt{n} ||\mathbf{E}'||_{\mathsf{op}} \\
&\overset{\text{(ii)}}{\geq} \frac{1}{||\boldsymbol{\lambda}||_1} - \sqrt{n} \cdot \frac{1}{||\boldsymbol{\lambda}||_1} \cdot \frac{1}{2\sqrt{n}} \\
&= \frac{1}{2||\boldsymbol{\lambda}||_1} > 0
\end{aligned}$$

for large enough n and with probability at least $\left(1 - \frac{2}{n}\right)$. Here, inequality (i) follows from the inequality $\mathbf{a}^\top \mathbf{M} \mathbf{b} \geq -||\mathbf{a}||_2 ||\mathbf{b}||_2 ||\mathbf{M}||_{\mathsf{op}}$, and inequality (ii) follows by substituting the upper bound we just derived on $||\mathbf{E}'||_{\mathsf{op}}$. This completes our proof for Part 1 of Theorem 11.

Next, we show that Equation (29) holds with high probability under the condition provided in Equation (12), which is the strictly sharper condition for the isotropic Gaussian case. For every $i \in [n]$, we denote $\mathbf{v}_i := \sqrt{n} \cdot Y_i \mathbf{e}_i$. We add and subtract terms to get

$$\begin{aligned}
Y_i \beta_i^* &= \frac{1}{\sqrt{n}} \mathbf{v}_i^\top \mathbf{A}^{-1} \mathbf{Y}_{\mathsf{train}} \\
&= \frac{1}{4\sqrt{n}} \left((\mathbf{v}_i + \mathbf{Y}_{\mathsf{train}})^\top \mathbf{A}^{-1} (\mathbf{v}_i + \mathbf{Y}_{\mathsf{train}}) - (\mathbf{v}_i - \mathbf{Y}_{\mathsf{train}})^\top \mathbf{A}^{-1} (\mathbf{v}_i - \mathbf{Y}_{\mathsf{train}}) \right).
\end{aligned}$$

We use the following technical lemma that shows concentration on quadratic forms of the inverse Wishart matrix $\mathbf{A}^{-1}$. From here on, we denote $d'(n) := (d - n + 1)$ for shorthand.

Lemma 21 *Let $\mathbf{A} \sim$ Wishart$(d, \mathbf{I}_n)$. For any vector $\mathbf{u} \in S^{n-1}$ and any $t > 0$, we have*

$$\Pr\left[\frac{1}{\mathbf{u}^\top \mathbf{A}^{-1}\mathbf{u}} > d'(n) + \sqrt{2t \cdot d'(n)} + 2t\right] \leq e^{-t}$$
$$\Pr\left[\frac{1}{\mathbf{u}^\top \mathbf{A}^{-1}\mathbf{u}} < d'(n) - \sqrt{2t \cdot d'(n)}\right] \leq e^{-t}.$$

provided that $d'(n) > 2\max\{t, 1\}$.

Lemma 21 is proved in Appendix F.2. Substituting the lower tail bound of Lemma 21 with $t := 2\ln n$ gives us

$$(\mathbf{v}_i + \mathbf{Y}_{\mathsf{train}})^\top \mathbf{A}^{-1}(\mathbf{v}_i + \mathbf{Y}_{\mathsf{train}}) \geq \frac{||\mathbf{v}_i + \mathbf{Y}_{\mathsf{train}}||_2^2}{d'(n) + \sqrt{4\ln n \cdot d'(n)} + 4\ln n}$$

with probability at least $\left(1 - \frac{1}{n^2}\right)$. Similarly, substituting the upper tail bound with $t := 2\ln n$ gives us

$$(\mathbf{v}_i - \mathbf{Y}_{\mathsf{train}})^\top \mathbf{A}^{-1}(\mathbf{v}_i - \mathbf{Y}_{\mathsf{train}}) \leq \frac{||\mathbf{v}_i - \mathbf{Y}_{\mathsf{train}}||_2^2}{d'(n) - \sqrt{4\ln n \cdot d'(n)}}$$

with probability at least $\left(1 - \frac{1}{n^2}\right)$. Noting that $||\mathbf{v}_i + \mathbf{Y}_{\mathsf{train}}||_2^2 = 2(n + \sqrt{n})$ and $||\mathbf{v}_i - \mathbf{Y}_{\mathsf{train}}||_2^2 = 2(n - \sqrt{n})$, we then get

$$\begin{aligned}
Y_i\beta_i^* &\propto \frac{n + \sqrt{n}}{d'(n) + \sqrt{4\ln n \cdot d'(n)} + 4\ln n} - \frac{n - \sqrt{n}}{d'(n) - \sqrt{4\ln n \cdot d'(n)}} \\
&= \frac{2\sqrt{n}d'(n) - 2n\sqrt{4\ln n \cdot d'(n)} - (4\ln n)(n - \sqrt{n})}{(d'(n) + \sqrt{4\ln n \cdot d'(n)})(d'(n) - \sqrt{4\ln n \cdot d'(n)})} \\
&> \frac{2\sqrt{n}d'(n) - 2n\sqrt{4\ln n \cdot d'(n)} - 4n \cdot \ln n}{(d'(n) + \sqrt{4\ln n \cdot d'(n)})(d'(n) - \sqrt{4\ln n \cdot d'(n)})} \\
&> 0
\end{aligned}$$

if we have

$$d'(n) > 9n\ln n \iff d > 9n\ln n + n - 1,$$

which is precisely the condition in Equation (12). Under this condition, we have proved that for any training data point corresponding to $i \in \{1, \ldots, n\}$, we have $Y_i\beta_i^* > 0$ with probability at least $\left(1 - \frac{2}{n^2}\right)$. Finally, applying the union bound on all n training data points gives us

$$Y_i\beta_i^* > 0 \text{ for all } i \in \{1, \ldots, n\}$$

with probability at least $\left(1 - \frac{2}{n}\right)$. This completes the proof of Theorem 11. ■

C.2 Implications of Theorem 11 for the bi-level ensemble

In this section, we provide the calculations that help us understand the ramifications of Theorem 11 — in particular, the condition in Equation (11) — for the bi-level ensemble (Definition 5). We reproduce Equation (11) below:

$$||\boldsymbol{\lambda}||_1 \geq 72 \left(||\boldsymbol{\lambda}||_2 \cdot n\sqrt{\ln n} + ||\boldsymbol{\lambda}||_\infty \cdot n\sqrt{n} \ln n + 1 \right).$$

We substitute the parameters of the bi-level ensemble into the left hand and right hand sides of the inequality. Recall that, by definition, we have $||\boldsymbol{\lambda}||_1 = d = n^p$ for the bi-level ensemble and so the left hand side is equal to n^p. On the other hand, for the right hand side, a simple calculation shows that

$$\begin{aligned} ||\boldsymbol{\lambda}||_2 &= \sqrt{s \cdot \frac{a^2 d^2}{s^2} + (d-s) \cdot \frac{(1-a)^2 d^2}{(d-s)^2}} \\ &\overset{\text{(i)}}{\asymp} \sqrt{\frac{a^2 d^2}{s} + d} \\ &= \sqrt{n^{2p-2q-r} + n^p} \\ &\asymp n^{\max\{p-q-\frac{r}{2}, \frac{p}{2}\}}, \end{aligned}$$

where the scaling in (i) follows because the bi-level ensemble defines $r < 1 < p$ and $q > 0$ (so $(1-a) \asymp 1$ and $(d-s) \asymp d$). Moreover, we have

$$||\boldsymbol{\lambda}||_\infty = \frac{ad}{s} = n^{p-q-r}.$$

Putting these together, the right hand side of Equation (11) scales as

$$\begin{aligned} &72 \left(||\boldsymbol{\lambda}||_2 \cdot n\sqrt{\ln n} + ||\boldsymbol{\lambda}||_\infty \cdot n\sqrt{n} \ln n + 1 \right) \\ &\asymp n^{\max\{p-q-\frac{r}{2}, \frac{p}{2}\}+1} \cdot \sqrt{\ln n} + n^{p+\frac{3}{2}-q-r} \cdot (\ln n) + 1, \end{aligned}$$

and so, for Equation (11) to hold, we get the following *sufficient* conditions on the parameters (p, q, r) of the bi-level ensemble for sufficiently large[21] n:

$$\begin{aligned} p > \frac{p}{2} + 1 &\implies p > 2 \\ p > p - q - \frac{r}{2} + 1 &\implies q > \left(1 - \frac{r}{2}\right) \\ p > p + \frac{3}{2} - q - r &\implies q > \left(\frac{3}{2} - r\right) \end{aligned}$$

Now, observe that $1 - \frac{r}{2} \leq \frac{3}{2} - r$ for all $0 \leq r \leq 1$, and so we get *sufficient conditions* as follows:

$$p > 2 \text{ and } q > \left(\frac{3}{2} - r\right),$$

These are precisely the conditions in Equation (13).

21. The reason for requiring sufficiently large n in these statements is the application of the $\asymp$ relation in multiple places. (Also note that Theorem 11 also required sufficiently large n.) Accordingly, we can also omit constants from consideration.

Appendix D. Proof of Theorem 13: Bounds on survival and contamination

In this section, we obtain a general, non-asymptotic characterization of classification (and regression) error by bounding survival and contamination terms. As described in Section 5.2, this is then plugged into the expressions in Proposition 17 to prove Theorem 13.

First, we define shorthand notation that is useful for this section, in addition to the notation already defined in Appendix B. For ease of notation, we denote the survival and contamination factors under the 1-sparse model for the case where we interpolate binary labels as

$$\mathsf{SU}_b(t) = \mathsf{SU}(\widehat{\boldsymbol{\alpha}}_{2,\mathsf{binary}}, t), \quad \mathsf{CN}_b(t) = \mathsf{CN}(\widehat{\boldsymbol{\alpha}}_{2,\mathsf{binary}}, t),$$

and for the case where we interpolate real output as

$$\mathsf{SU}_r(t) = \mathsf{SU}(\widehat{\boldsymbol{\alpha}}_{2,\mathsf{real}}, t), \quad \mathsf{CN}_r(t) = \mathsf{CN}(\widehat{\boldsymbol{\alpha}}_{2,\mathsf{real}}, t).$$

Finally, for a given index $t \in \{1, \ldots, d\}$, we denote as shorthand $\mathbf{z}_t := \mathbf{Z}_{\mathsf{train}}$. It is easy to verify that $\mathbf{z}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ under the 1-sparse Assumption 1. We also denote $\mathbf{y}_t := \mathbf{Y}_{\mathsf{train}}$. Recall that we consider the possibility of label noise probability equal to ν^*: from the generative model defined in Equation (1), we have

$$y_{t,i} = \begin{cases} \mathsf{sgn}(z_{t,i}) \text{ with probability } (1-\nu^*) \\ -\mathsf{sgn}(z_{t,i}) \text{ with probability } \nu^*. \end{cases} \tag{32}$$

for every $i \in \{1, \ldots, n\}$. Finally, for a given positive semi-definite matrix $\mathbf{M} \in \mathbb{R}^{d \times d}$ and a given index
$k \in \{0, \ldots, (d-1)\}$, we define the *effective rank*

$$r_k(\mathbf{M}) := \frac{\sum_{\ell > k} \mu_\ell(\mathbf{M})}{\mu_{k+1}(\mathbf{M})}.$$

Recall that this is the precisely the definition of the first effective rank in Bartlett et al. (2020), which dictates the contribution of pure signal to regression test error incurred by the minimum-ℓ_2-norm interpolation.

D.1 Bounds on survival and contamination

The notions of survival and contamination were first introduced in Muthukumar et al. (2020), and characterized there with equality for Fourier featurization on regularly spaced training data. Here, we characterize these quantities for Gaussian features. We state our upper and lower bounds on survival and contamination respectively for two cases — when the output being interpolated is binary, and when the output being interpolated is real. We start with upper and lower bounds on the survival factor.

Theorem 22 (Upper and lower bounds on survival) *There exist universal positive constants (b, b_2, c, c_3, c_4) (that do not depend on parameters $(n, d, k, \boldsymbol{\Sigma})$) such that if $r_k(\boldsymbol{\Sigma}) \geq bn$ and $r_k(\boldsymbol{\Sigma}_{-t}) \geq b_2 n$, we have the following characterizations of the survival factor for any $k \geq t$:*

1. ***Interpolation of binary labels:*** *The minimum-ℓ_2-norm interpolation of binary labels, i.e. $\widehat{\boldsymbol{\alpha}}_{2,\mathsf{binary}}$, satisfies each of*

$$\mathsf{SU}_b(t) \geq \sqrt{\frac{2}{\pi}} \cdot (1-2\nu^*) \cdot \frac{\lambda_t\left(\frac{(n-k)}{c\widetilde{\lambda}_{k+1} r_k(\boldsymbol{\Sigma}_{-t})} - \frac{c_3 n^{3/4}}{\lambda_{k+1} r_k(\boldsymbol{\Sigma})}\right)}{1+\lambda_t\left(\frac{cn}{\widetilde{\lambda}_{k+1} r_k(\boldsymbol{\Sigma}_{-t})} + \frac{c_4 n^{3/4}}{\lambda_{k+1} r_k(\boldsymbol{\Sigma})}\right)}, \text{ and} \tag{33a}$$

$$\mathsf{SU}_b(t) \leq \sqrt{\frac{2}{\pi}} \cdot (1-2\nu^*) \cdot \frac{\lambda_t\left(\frac{cn}{\widetilde{\lambda}_{k+1} r_k(\boldsymbol{\Sigma}_{-t})} + \frac{c_3 n^{3/4}}{\lambda_{k+1} r_k(\boldsymbol{\Sigma})}\right)}{1+\lambda_t\left(\frac{(n-k)}{c\widetilde{\lambda}_{k+1} r_k(\boldsymbol{\Sigma}_{-t})} - \frac{c_4 n^{3/4}}{\lambda_{k+1} r_k(\boldsymbol{\Sigma})}\right)} \tag{33b}$$

with probability at least $(1-3e^{-\sqrt{n}}-2e^{-\frac{n}{c}})$ over the randomness in the training data $\{X_i, Y_i\}_{i=1}^n$.

2. ***Interpolation of real output:*** *The minimum-ℓ_2-norm interpolation of real output, i.e. $\widehat{\boldsymbol{\alpha}}_{2,\mathsf{real}}$, satisfies each of*

$$\mathsf{SU}_r(t) \geq \frac{1}{1+\frac{1}{\lambda_t\left(\frac{(n-k)}{c\widetilde{\lambda}_{k+1} r_k(\boldsymbol{\Sigma}_{-t})} - \frac{c_4 n^{\frac{3}{4}}}{\lambda_{k+1} r_k((\boldsymbol{\Sigma})}\right)}}, \text{ and} \tag{34a}$$

$$\mathsf{SU}_r(t) \leq \frac{1}{1+\frac{1}{\lambda_t\left(\frac{cn}{\widetilde{\lambda}_{k+1} r_k(\boldsymbol{\Sigma}_{-t})} + \frac{c_4 n^{\frac{3}{4}}}{\lambda_{k+1} r_k((\boldsymbol{\Sigma})}\right)}} \tag{34b}$$

with probability at least $(1-2e^{-\sqrt{n}}-2e^{-\frac{n}{c}})$ over the randomness in the training data $\{X_i, Y_i\}_{i=1}^n$.

We will see subsequently (in Appendix E) that the survival bounds, whether binary labels or real output are interpolated, are matching in their dependence on n up to constants. We now state our characterization of the contamination factor.

Theorem 23 (Upper and lower bounds on contamination) *There exist universal positive constants $b_2, c_5, c_6, c_7, c_8, c_9$ (that do not depend on parameters $(n, d, k, \boldsymbol{\Sigma})$) such that if $0 \leq k \leq n/c_5$ and $r_k(\boldsymbol{\Sigma}_{-t}) \geq b_2$, the following characterizations of the contamination factor hold for any choice of $\ell \leq k$:*

1. ***Interpolation of binary labels:*** *Provided that $n \geq c_6$, the minimum-ℓ_2-norm interpolation of binary labels, i.e. $\widehat{\boldsymbol{\alpha}}_{2,\mathsf{binary}}$, satisfies each of*

$$\mathsf{CN}_b(t) \leq c_7 \cdot \sqrt{\left(\frac{\ell}{n} + n \cdot \frac{\sum_{j>\ell} \widetilde{\lambda}_j^2}{\left(\sum_{j>k} \widetilde{\lambda}_j\right)^2}\right) \cdot \ln n \cdot (1+\mathsf{SU}_b(t)^2)}, \text{ and} \tag{35a}$$

$$\mathsf{CN}_b(t) \geq \sqrt{n} \cdot \frac{\sqrt{r_k\left(\boldsymbol{\Sigma}_{-t}^2\right) \cdot \widetilde{\lambda}_{k+1}^2}}{c_9\left(\sum_{j=1}^{d} \lambda_j + \lambda_1 n\right)} \tag{35b}$$

almost surely for any realization of the random quantity $\mathsf{SU}_b(t)$*, and with probability at least* $\left(1-\frac{3}{n}\right)$ *and* $(1-2e^{-\frac{n}{c_8}})$ *respectively over the randomness in the training data* $\{X_i, Y_i\}_{i=1}^n$.

2. ***Interpolation of real output:*** *Provided that* $n \geq c_6$*, the minimum-*ℓ_2*-norm interpolation of real output, i.e.* $\widehat{\boldsymbol{\alpha}}_{2,\mathsf{real}}$*, satisfies each of*

$$\mathsf{CN}_r(t) \leq c_7 \left|1-\mathsf{SU}_r(t)\right| \cdot \sqrt{\left(\frac{l}{n} + n \cdot \frac{\sum_{j>l} \widetilde{\lambda}_j^2}{\left(\sum_{j>k} \widetilde{\lambda}_j\right)^2}\right) \cdot \ln n}, \text{ and} \tag{36a}$$

$$\mathsf{CN}_r(t) \geq \sqrt{n(1-\delta)} \cdot \frac{\sqrt{r_k\left(\boldsymbol{\Sigma}_{-t}^2\right) \cdot \widetilde{\lambda}_{k+1}^2}}{c_9\left(\sum_{j=1}^d \lambda_j + \lambda_1 n\right)} \tag{36b}$$

almost surely for any realization of the random quantity $\mathsf{SU}_b(t)$*, and with probability at least* $\left(1-\frac{2}{n}\right)$ *and* $(1-2e^{-\frac{n}{c_8}}-e^{-n\delta^2})$ *respectively over the randomness in the training data* $\{X_i, Y_i\}_{i=1}^n$.

Observe that the high-probability characterizations of contamination in Theorem 23 themselves hold almost surely for every realization of the respective survival factors for binary and real interpolation, which are random variables. In Appendix E, these expressions will be used together (with a simple union bound) with the matching high-probability characterization of survival factor in Theorem 22. Unlike for the case of survival, the upper and lower bounds for contamination are not necessarily matching — however, as we will see in Appendix E, they turn out to match for all parameterizations of the bi-level ensemble.

As a final remark, in both theorem statements, the only randomness over which all probabilities are taken is solely in the training data $\{X_i, Y_i\}_{i=1}^n$. Further, all universal positive constants are taken to be independent of the parameters $(n, d, k, \boldsymbol{\Sigma})$, which entirely describe the problem. In the proofs of Theorems 22 and 23, we will follow these conventions unless specified otherwise.

D.2 Background lemmas

We begin our proofs of Theorems 22 and 23 by stating lemmas that serve as background for our analysis. The first lemma is from Bartlett et al. (2020).

Lemma 24 ***(Concentration of eigenvalues, Lemmas 9 and 10 in Bartlett et al., 2020)*** *There exist universal positive constants* (b, c) *such that:*

1. *For any* $k \geq 0$ *such that* $r_k(\boldsymbol{\Sigma}) \geq bn$*, we have*

$$\frac{1}{c}\lambda_{k+1} r_k(\boldsymbol{\Sigma}) \leq \mu_n\left(\mathbf{A}\right) \leq \mu_1\left(\mathbf{A}\right) \leq c\left(\sum_{j=1}^d \lambda_j + \lambda_1 n\right) \text{ and} \tag{37}$$

$$\mu_{k+1}(\mathbf{A}) \leq c\lambda_{k+1} r_k(\boldsymbol{\Sigma}) \tag{38}$$

with probability at least $(1-2e^{-\frac{n}{c}})$ *over the random matrix* $\mathbf{A}$.

2. *For any $k \geq t$ such that $r_k(\mathbf{\Sigma}) \geq bn$, we have*

$$\frac{1}{c}\lambda_{k+1} r_k(\mathbf{\Sigma}) \leq \mu_n\left(\mathbf{A}_{-t}\right) \leq \mu_1\left(\mathbf{A}_{-t}\right) \leq c\left(\sum_{j=1}^{d} \lambda_j + \lambda_1 n\right) \tag{39}$$

with probability at least $(1 - 2e^{-\frac{n}{c}})$ over the random matrix $\mathbf{A}_{-t}$.

Further, as corollaries to the above, we have the following statements:

1. *For any $k \geq 0$ such that $r_k(\mathbf{\Sigma}) \geq bn$, we have*

$$\frac{1}{c\left(\sum_{j=1}^{d} \lambda_j + \lambda_1 n\right)} \leq \mu_n\left(\mathbf{A}^{-1}\right) \leq \mu_1\left(\mathbf{A}^{-1}\right) \leq \frac{c}{\lambda_{k+1} r_k(\mathbf{\Sigma})} \tag{40}$$

with probability at least $(1 - 2e^{-\frac{n}{c}})$ over the random matrix $\mathbf{A}$.

2. *For any $k \geq t$ such that $r_k(\mathbf{\Sigma}) \geq bn$, we have*

$$\frac{1}{c\left(\sum_{j=1}^{d} \lambda_j + \lambda_1 n\right)} \leq \mu_n\left(\mathbf{A}_{-t}^{-1}\right) \leq \mu_1\left(\mathbf{A}_{-t}^{-1}\right) \leq \frac{c}{\lambda_{k+1} r_k(\mathbf{\Sigma})} \tag{41}$$

with probability at least $(1 - 2e^{-\frac{n}{c}})$ over the random matrix $\mathbf{A}_{-t}$.

Note that using Equation (28) to express $\mathbf{A}_{-t}$, we can rewrite the bounds in the above lemma in terms of the quantities $\mathbf{\Sigma}_{-t}$ and $\widetilde{\lambda}_j$. In particular, it follows that each of

$$\frac{1}{c}\widetilde{\lambda}_{k+1} r_k(\mathbf{\Sigma}_{-t}) \leq \mu_n\left(\mathbf{A}_{-t}\right) \leq \mu_1\left(\mathbf{A}_{-t}\right) \leq c\left(\sum_{j=1}^{d-1} \widetilde{\lambda}_j + \widetilde{\lambda}_1 n\right) \text{ and} \tag{42a}$$

$$\frac{1}{c\left(\sum_{j=1}^{d-1} \widetilde{\lambda}_j + \widetilde{\lambda}_1 n\right)} \leq \mu_n\left(\mathbf{A}_{-t}^{-1}\right) \leq \mu_1\left(\mathbf{A}_{-t}^{-1}\right) \leq \frac{c}{\widetilde{\lambda}_{k+1} r_k(\mathbf{\Sigma}_{-t})}. \tag{42b}$$

holds with probability at least $(1 - 2e^{-\frac{n}{c}})$. We will also apply Equation (38) with $\mathbf{A}_{-t}$ instead of $\mathbf{A}$, and use the corresponding condition $r_k(\mathbf{\Sigma}) \geq b_2 n$.
The next lemma is the Hanson-Wright inequality, which shows that the quadratic form of a (sub)-Gaussian random vector concentrates around its expectation.

Lemma 25 ***(Hanson-Wright inequality, Rudelson and Vershynin, 2013)*** *Let $\mathbf{z}$ be a random vector composed of i.i.d. random variables that are zero mean and sub-Gaussian with parameter at most 1. Then, there exists universal constant $c > 0$ such that for any positive semi-definite matrix $\mathbf{M}$ and for every $t \geq 0$, we have*

$$\Pr\left[|\mathbf{z}^\top \mathbf{M}\mathbf{z} - \mathbb{E}[\mathbf{z}^\top \mathbf{M}\mathbf{z}]| > t\right] \leq 2\exp\left\{-c\min\left\{\frac{t^2}{||\mathbf{M}||_{\mathsf{F}}^2}, \frac{t}{||\mathbf{M}||_{\mathsf{op}}}\right\}\right\}.$$

We will apply this inequality in two ways. First, we will note that $||\mathbf{M}||_{\mathsf{F}}^2 \leq n||\mathbf{M}||_{\mathsf{op}}^2$ and substitute $t := c_1||\mathbf{M}||_{\mathsf{op}} \cdot n^{3/4}$ (where $c_1^2 = \frac{1}{c}$) to get

$$|\mathbf{z}^\top \mathbf{M}\mathbf{z} - \mathbb{E}[\mathbf{z}^\top \mathbf{M}\mathbf{z}]| \leq c_1||\mathbf{M}||_{\mathsf{op}} \cdot n^{3/4} \tag{43}$$

with probability at least $(1 - 2e^{-\sqrt{n}})$. Second, we will note that $||\mathbf{M}||_{\mathsf{op}} \leq \mathsf{tr}(\mathbf{M})$ and moreover, $||\mathbf{M}||_{\mathsf{F}}^2 = \mathsf{tr}(\mathbf{M}^2) \leq (\mathsf{tr}(\mathbf{M}))^2$. Then, substituting $t := \frac{1}{c} \cdot \mathsf{tr}(\mathbf{M}) \cdot (\ln n)$, we get

$$\mathbf{z}^\top \mathbf{M}\mathbf{z} \leq \mathbb{E}[\mathbf{z}^\top \mathbf{M}\mathbf{z}] + \frac{1}{c} \cdot \mathsf{tr}(\mathbf{M}) \cdot (\ln n) \leq \left(1 + \frac{1}{c}\right) \cdot \mathsf{tr}(\mathbf{M}) \cdot (\ln n) \tag{44}$$

with probability at least $(1 - \frac{1}{n})$. Finally, note that all probabilities are only over the random vector $\mathbf{z}$. We will frequently apply Lemma 25 as a high-probability statement conditioned on the realization of a *random*, almost surely positive semi-definite matrix $\mathbf{M}$ which is independent of $\mathbf{z}$.

Finally, the following lemma bounds the squared norm of a Gaussian random vector by a standard tail bound on chi-squared random variables (for e.g. see Wainwright, 2019, Chapter 2), stated for completeness.

Lemma 26 *Let $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$. Then, for any $\delta \in (0, 1)$, we have*

$$n(1 - \delta) \leq \|\mathbf{z}\|_2^2 \leq n(1 + \delta) \tag{45}$$

with probability at least $(1 - 2e^{-n\delta^2})$.

D.3 Proof of Theorem 22

We first prove Theorem 22, i.e. upper and lower bounds on survival when binary labels *or* real output are interpolated. We start with the slightly more difficult case of interpolation of binary labels (Equations (33a) and (33b)).

D.3.1 Interpolation of binary labels

Recall that, by Assumption 1, we have $\alpha_t^* = \frac{1}{\sqrt{\lambda_t}}$. A standard argument based on Moore-Penrose pseudoinverse calculations shows that $\widehat{\boldsymbol{\alpha}}_{2,\mathsf{binary}} = \boldsymbol{\Phi}_{\mathsf{train}}^\top (\boldsymbol{\Phi}_{\mathsf{train}} \boldsymbol{\Phi}_{\mathsf{train}}^\top)^{-1} \mathbf{Y}_{\mathsf{train}}$. We get

$$\begin{aligned}
\mathsf{SU}_b(t) &= \frac{\widehat{\alpha}_{t,2,\mathsf{binary}}}{\alpha_t^*} \\
&= \sqrt{\lambda_t}\widehat{\alpha}_{t,2,\mathsf{binary}} \\
&= \sqrt{\lambda_t}\mathbf{e}_t^\top \boldsymbol{\Phi}_{\mathsf{train}}^\top (\boldsymbol{\Phi}_{\mathsf{train}} \boldsymbol{\Phi}_{\mathsf{train}}^\top)^{-1} \mathbf{Y}_{\mathsf{train}} \\
&= \lambda_t \mathbf{z}_t^\top \mathbf{A}^{-1} \mathbf{y}_t,
\end{aligned}$$

where $\mathbf{z}_t, \mathbf{y}_t$ are as defined at the beginning of Appendix D, and $\mathbf{A}$ is the Gram matrix defined in Appendix B. Next, we use the Sherman-Morrison-Woodbury identity to get

$$\begin{aligned}
\mathbf{A}^{-1} &= (\lambda_t \mathbf{z}_t \mathbf{z}_t^\top + \mathbf{A}_{-t})^{-1} \\
&= \mathbf{A}_{-t}^{-1} - \frac{\lambda_t \mathbf{A}_{-t}^{-1} \mathbf{z}_t \mathbf{z}_t^\top \mathbf{A}_{-t}^{-1}}{1 + \lambda_t \mathbf{z}_t^\top \mathbf{A}_{-t}^{-1} \mathbf{z}_t}.
\end{aligned} \tag{46}$$

Using this, we obtain

$$\mathsf{SU}_b(t) = \frac{\lambda_t \mathbf{z}_t^\top \mathbf{A}_{-t}^{-1} \mathbf{y}_t}{1 + \lambda_t \mathbf{z}_t^\top \mathbf{A}_{-t}^{-1} \mathbf{z}_t}. \tag{47}$$

Adding and subtracting terms to the numerator, we get

$$\mathbf{z}_t^\top \mathbf{A}_{-t}^{-1} \mathbf{y}_t = \frac{1}{4}\left((\mathbf{z}_t + \mathbf{y}_t)^\top \mathbf{A}_{-t}^{-1} (\mathbf{z}_t + \mathbf{y}_t) - (\mathbf{z}_t - \mathbf{y}_t)^\top \mathbf{A}_{-t}^{-1} (\mathbf{z}_t - \mathbf{y}_t) \right).$$

Because of the "leave-one-out" property, note that $\mathbf{A}_{-t}^{-1} \perp \{\mathbf{z}_t, \mathbf{y}_t\}$. Also note that $\mathbf{A}_{-t}^{-1}$ is almost surely positive semidefinite. Thus, we can upper *and* lower bound the numerator of Equation (47) around its expectation using the Hanson-Wright inequality. First, we calculate the conditional expectation:

$$\begin{aligned}
\mathbb{E}\left[\mathbf{z}_t^\top \mathbf{A}_{-t}^{-1} \mathbf{y}_t \middle| \mathbf{A}_{-t}^{-1} \right] &= \mathbb{E}\left[\mathsf{tr}(\mathbf{A}_{-t}^{-1} \mathbf{y}_t \mathbf{z}_t^\top) \middle| \mathbf{A}_{-t}^{-1} \right] \\
&= \mathsf{tr}\left(\mathbf{A}_{-t}^{-1} \cdot \mathbb{E}\left[\mathbf{y}_t \mathbf{z}_t^\top \right] \right).
\end{aligned}$$

Recalling the expression for $\mathbf{y}_t$ from Equation (32), a simple calculation yields that

$$\begin{aligned}
\mathbb{E}\left[\mathbf{y}_t \mathbf{z}_t^\top \right] &= \mathbb{E}\left[y_{t,1} z_{t,1}^\top \right] \cdot \mathbf{I}_n \\
&= \left((1-\nu^*) \mathbb{E}\left[\mathsf{sgn}(z_{t,1}) z_{t,1}^\top \right] + \nu^* \mathbb{E}\left[-\mathsf{sgn}(z_{t,1}) z_{t,1}^\top \right] \right) \cdot \mathbf{I}_n \\
&= (1 - 2\nu^*) \mathbb{E}\left[\mathsf{sgn}(z_{t,1}) z_{t,1}^\top \right] \cdot \mathbf{I}_n \\
&= (1 - 2\nu^*) \cdot \sqrt{\frac{2}{\pi}} \cdot \mathbf{I}_n,
\end{aligned}$$

where the last step follows because $z_{t,1} \sim \mathcal{N}(0,1)$.

Now, we apply Equation (43) (the Hanson-Wright inequality) *almost surely* for every realization of the random matrix $\mathbf{A}_{-t}^{-1}$, and simultaneously to the quadratic forms $(\mathbf{z}_t + \mathbf{y}_t)^\top \mathbf{A}_{-t}^{-1} (\mathbf{z}_t + \mathbf{y}_t)$ and $(\mathbf{z}_t - \mathbf{y}_t)^\top \mathbf{A}_{-t}^{-1} (\mathbf{z}_t - \mathbf{y}_t)$. Thus, we have each of

$$\begin{aligned}
\mathbf{z}_t^\top \mathbf{A}_{-t}^{-1} \mathbf{y}_t &\geq \left((1-2\nu^*) \sqrt{\frac{2}{\pi}} \mathsf{tr}(\mathbf{A}_{-t}^{-1}) - 2c_1 ||\mathbf{A}_{-t}^{-1}||_{\mathsf{op}} \cdot n^{3/4} \right) \text{ and} \\
\mathbf{z}_t^\top \mathbf{A}_{-t}^{-1} \mathbf{y}_t &\leq \left((1-2\nu^*) \sqrt{\frac{2}{\pi}} \mathsf{tr}(\mathbf{A}_{-t}^{-1}) + 2c_1 ||\mathbf{A}_{-t}^{-1}||_{\mathsf{op}} \cdot n^{3/4} \right)
\end{aligned}$$

with probability at least $(1 - 2e^{-\sqrt{n}})$ over the randomness in $\{\mathbf{z}_t, \mathbf{y}_t\}$. Similarly, to bound the the denominator, we have each of

$$\mathbf{z}_t^\top \mathbf{A}_{-t}^{-1} \mathbf{z}_t \geq \mathsf{tr}(\mathbf{A}_{-t}^{-1}) - c_1 ||\mathbf{A}_{-t}^{-1}||_{\mathsf{op}} \cdot n^{3/4} \text{ and} \tag{48a}$$

$$\mathbf{z}_t^\top \mathbf{A}_{-t}^{-1} \mathbf{z}_t \leq \mathsf{tr}(\mathbf{A}_{-t}^{-1}) + c_1 ||\mathbf{A}_{-t}^{-1}||_{\mathsf{op}} \cdot n^{3/4} \tag{48b}$$

with probability at least $(1 - e^{-\sqrt{n}})$ over the randomness in $\{\mathbf{z}_t, \mathbf{y}_t\}$. Substituting these bounds into Equation (47), we get each of

$$\mathsf{SU}_b(t) \geq \frac{\lambda_t \cdot \left(\sqrt{\frac{2}{\pi}}(1-2\nu^*)\mathsf{tr}(\mathbf{A}_{-t}^{-1}) - 2c_1||\mathbf{A}_{-t}^{-1}||_{\mathsf{op}} \cdot n^{3/4}\right)}{1 + \lambda_t \left(\mathsf{tr}(\mathbf{A}_{-t}^{-1}) + c_1||\mathbf{A}_{-t}^{-1}||_{\mathsf{op}} \cdot n^{3/4}\right)} \text{ and}$$

$$\mathsf{SU}_b(t) \leq \frac{\lambda_t \cdot \left(\sqrt{\frac{2}{\pi}}(1-2\nu^*)\mathsf{tr}(\mathbf{A}_{-t}^{-1}) + 2c_1||\mathbf{A}_{-t}^{-1}||_{\mathsf{op}} \cdot n^{3/4}\right)}{1 + \lambda_t \left(\mathsf{tr}(\mathbf{A}_{-t}^{-1}) - c_1||\mathbf{A}_{-t}^{-1}||_{\mathsf{op}} \cdot n^{3/4}\right)},$$

with probability at least $(1 - 3e^{-\sqrt{n}})$ over the randomness in $\{\mathbf{z}_t, \mathbf{y}_t\}$. It remains to obtain high-probability bounds on the random quantities $\mathsf{tr}(\mathbf{A}_{-t}^{-1})$ and $||\mathbf{A}_{-t}^{-1}||_{\mathsf{op}}$. Note that we need both lower bounds and upper bounds on the quantity $\mathsf{tr}(\mathbf{A}_{-t}^{-1})$, but we only need an upper bound on the quantity $||\mathbf{A}_{-t}^{-1}||_{\mathsf{op}}$.

We assume that we can choose $k \geq t$ such that $r_k(\mathbf{\Sigma}) \geq bn$ and $r_k(\mathbf{\Sigma}_{-t}) \geq b_2 n$ for universal positive constants (b, b_2). Consider any such choice of k (which in general could depend on (n, d)). First, we use Equation (41) from Lemma 24 to upper bound the quantity $||\mathbf{A}_{-t}^{-1}||_{\mathsf{op}}$ as

$$||\mathbf{A}_{-t}^{-1}||_{\mathsf{op}} = \mu_1(\mathbf{A}_{-t}^{-1}) \leq \frac{c}{\lambda_{k+1} r_k(\mathbf{\Sigma})} \tag{49}$$

with probability at least $(1 - e^{-\frac{n}{c}})$ over the random matrix $\mathbf{A}$. Next, we turn to the quantity $\mathsf{tr}(\mathbf{A}_{-t}^{-1})$. To lower bound this quantity, we notice that

$$\begin{aligned}
\mathsf{tr}(\mathbf{A}_{-t}^{-1}) &= \sum_{j=1}^{n} \frac{1}{\mu_j(\mathbf{A}_{-t})} \\
&\geq \sum_{j=k}^{n} \frac{1}{\mu_j(\mathbf{A}_{-t})} \\
&\geq \frac{(n-k)}{\mu_{k+1}(\mathbf{A}_{-t})}.
\end{aligned}$$

Now, from Equation (38) in Lemma 24 applied with $\mathbf{A}_{-t}$, we have

$$\mu_{k+1}(\mathbf{A}_{-t}) \leq c\widetilde{\lambda}_{k+1} r_k(\mathbf{\Sigma}_{-t})$$

with probability at least $(1 - e^{-\frac{n}{c}})$ provided that $r_k(\mathbf{\Sigma}_{-t}) \geq b_2 n$. This gives us:

$$\mathsf{tr}(\mathbf{A}_{-t}^{-1}) \geq \frac{(n-k)}{c\widetilde{\lambda}_{k+1} r_k(\mathbf{\Sigma}_{-t})}. \tag{50}$$

with probability at least $(1 - e^{-\frac{n}{c}})$. On the other hand, the upper bound on the trace follows simply by

$$\begin{aligned}
\mathsf{tr}(\mathbf{A}_{-t}^{-1}) &\leq \frac{n}{\mu_n(\mathbf{A}_{-t})} \\
&\leq \frac{cn}{\widetilde{\lambda}_{k+1} r_k(\mathbf{\Sigma}_{-t})},
\end{aligned} \tag{51}$$

where the last inequality substitutes Equation (42a), which again holds with probability at least $(1 - e^{-\frac{n}{c}})$. Noting that the upper bound on $\mathsf{SU}_b(t)$ is monotonically increasing in both $\mathsf{tr}(\mathbf{A}_{-t}^{-1})$ and $||\mathbf{A}_{-t}^{-1}||_{\mathsf{op}}$, and the lower bound on $\mathsf{SU}_b(t)$ is monotonically increasing in $\mathsf{tr}(\mathbf{A}_{-t}^{-1})$ but decreasing in $||\mathbf{A}_{-t}^{-1}||_{\mathsf{op}}$, we can substitute the above bounds on these quantities. This completes our characterization of survival when binary labels are interpolated, with the probability of this characterization lower bounded by taking a union bound over the complement of all the above events. After taking this union bound, the probability of each of the lower bound (Equation (33a)) and upper bound (Equation (33b)) holding is *at least* $(1 - 3e^{-\sqrt{n}} - 2e^{-\frac{n}{c}})$.

D.3.2 Interpolation of real output

For completeness, we also include the proof of Theorem 22 for the simpler case of interpolation of real-valued output (Equations (34a) and (34b)). By the same standard argument, we can characterize the minimum-ℓ_2-norm interpolator of real output as $\widehat{\boldsymbol{\alpha}}_{2,\mathsf{real}} = \boldsymbol{\Phi}_{\mathsf{train}}^\top(\boldsymbol{\Phi}_{\mathsf{train}}\boldsymbol{\Phi}_{\mathsf{train}}^\top)^{-1}\mathbf{Z}_{\mathsf{train}}$. By a similar argument to the case of binary labels, we have

$$\begin{aligned}\mathsf{SU}_r(t) &= \sqrt{\lambda_t}\widehat{\alpha}_t \\ &= \sqrt{\lambda_t}\mathbf{e}_t^\top\boldsymbol{\Phi}_{\mathsf{train}}^\top(\boldsymbol{\Phi}_{\mathsf{train}}\boldsymbol{\Phi}_{\mathsf{train}}^\top)^{-1}\mathbf{Z}_{\mathsf{train}} \\ &= \lambda_t\mathbf{z}_t^\top\mathbf{A}^{-1}\mathbf{z}_t.\end{aligned}$$

Again, using the Sherman-Morrison-Woodbury identity, we have

$$\mathbf{A}^{-1} = \mathbf{A}_{-t}^{-1} - \frac{\lambda_t\mathbf{A}_{-t}^{-1}\mathbf{z}_t\mathbf{z}_t^\top\mathbf{A}_{-t}^{-1}}{1 + \lambda_t\mathbf{z}_t^\top\mathbf{A}_{-t}^{-1}\mathbf{z}_t},$$

which gives us

$$\begin{aligned}\mathsf{SU}_r(t) &= \frac{\lambda_t\mathbf{z}_t^\top\mathbf{A}_{-t}^{-1}\mathbf{z}_t}{1 + \lambda_t\mathbf{z}_t^\top\mathbf{A}_{-t}^{-1}\mathbf{z}_t} \\ &= \frac{1}{1 + \frac{1}{\lambda_t\mathbf{z}_t^\top\mathbf{A}_{-t}^{-1}\mathbf{z}_t}}. \end{aligned} \tag{52}$$

From Equations (48a) and (48b) above, the following statements each hold with probability at least $(1 - e^{-\sqrt{n}})$ over the randomness in $\mathbf{z}_t$ and for every realization of the random matrix $\mathbf{A}_{-t}^{-1}$:

$$\begin{aligned}\mathbf{z}_t^\top\mathbf{A}_{-t}^{-1}\mathbf{z}_t &\geq \mathsf{tr}(\mathbf{A}_{-t}^{-1}) - c_2||\mathbf{A}_{-t}^{-1}||_{\mathsf{op}} \cdot n^{3/4} \text{ and} \\ \mathbf{z}_t^\top\mathbf{A}_{-t}^{-1}\mathbf{z}_t &\leq \mathsf{tr}(\mathbf{A}_{-t}^{-1}) + c_2||\mathbf{A}_{-t}^{-1}||_{\mathsf{op}} \cdot n^{3/4}.\end{aligned}$$

Here, c_2 is a universal positive constant.

Observe that the right hand side of Equation (52) is increasing in the quantity $\mathbf{z}_t^\top\mathbf{A}_{-t}^{-1}\mathbf{z}_t$. Thus, substituting the lower bound for $\mathsf{tr}(\mathbf{A}_{-t}^{-1})$ from Equation (50) and the upper bound for $||\mathbf{A}_{-t}^{-1}||_{\mathsf{op}}$ from Equation (49) lower bounds the quantity $\mathbf{z}_t^\top\mathbf{A}_{-t}^{-1}\mathbf{z}_t$, yielding the lower bound for $\mathsf{SU}_r(t)$. Similarly, substituting the upper bound for $\mathsf{tr}(\mathbf{A}_{-t}^{-1})$ from Equation (51) and the upper bound for $||\mathbf{A}_{-t}^{-1}||_{\mathsf{op}}$ from Equation (49) upper bounds the quantity $\mathbf{z}_t^\top\mathbf{A}_{-t}^{-1}\mathbf{z}_t$,

yielding the upper bound for $\mathsf{SU}_r(t)$. This completes the proof of Theorem 22. Again, a simple application of the union bound shows that each of the lower bound (Equation (34a)) and the upper bound (Equation (34b)) hold with probability at least $(1 - 2e^{-\sqrt{n}} - 2e^{-\frac{n}{c}})$.

D.4 Proof of Theorem 23

We next prove Theorem 23, i.e. upper and lower bounds on contamination, for the cases of interpolating binary labels and real output. Since the contamination factor is intricately related to the contribution of additive noise to regression test error, the proof primarily consists of refinements of the arguments in Bartlett et al. (2020).

D.4.1 Interpolation of binary labels

We start with a useful set of expressions for the contamination factor in the following lemma. The proof of this lemma is contained in Appendix F.3.

Lemma 27 *The contamination of the minimum-ℓ_2-norm interpolation of binary labels, denoted by $\widehat{\boldsymbol{\alpha}}_{2,\mathsf{binary}}$, can be written in the following two forms:*

$$\mathsf{CN}_b(t) = \sqrt{\mathbf{y}_t^\top \mathbf{C} \mathbf{y}_t}, \tag{53a}$$

$$= \sqrt{\widetilde{\mathbf{y}}_t^{\,\top} \widetilde{\mathbf{C}} \widetilde{\mathbf{y}}_t}, \tag{53b}$$

where we denote

$$\widetilde{\mathbf{y}}_t := \mathbf{y}_t - \mathsf{SU}_b(t)\mathbf{z}_t \ ,$$

$$\mathbf{C} := \mathbf{A}^{-1} \left(\sum_{j=1, j\neq t}^{d} \lambda_j^2 \mathbf{z}_j \mathbf{z}_j^\top \right) \mathbf{A}^{-1} \ , \text{ and}$$

$$\widetilde{\mathbf{C}} := \mathbf{A}_{-t}^{-1} \left(\sum_{j=1, j\neq t}^{d} \lambda_j^2 \mathbf{z}_j \mathbf{z}_j^\top \right) \mathbf{A}_{-t}^{-1}.$$

We will use the expression in Equation (53b) to prove an upper bound on contamination, and the expression in Equation (53a) for the lower bound.

D.4.2 Upper bound on $\mathsf{CN}_b(t)$

We start with the proof for the upper bound on contamination for interpolation of binary labels (Equation (35a)). From Equation (53b) in Lemma 27, we have $\mathsf{CN}_b^2(t) = \widetilde{\mathbf{y}}_t^{\,\top} \widetilde{\mathbf{C}} \widetilde{\mathbf{y}}_t$. Note that by construction, $\widetilde{\mathbf{C}}$ has no dependence on $\{\mathbf{z}_t, \mathbf{y}_t\}$ and thus $\widetilde{\mathbf{C}} \perp \widetilde{\mathbf{y}}_t$. The next lemma upper bounds the term $\widetilde{\mathbf{y}}_t^{\,\top} \widetilde{\mathbf{C}} \widetilde{\mathbf{y}}_t$ in terms of $\mathsf{tr}(\widetilde{\mathbf{C}})$ and is proved in Appendix F.4.

Lemma 28 *There exists universal positive constant c_6 such that when $n \geq c_6$, we have*

$$\widetilde{\mathbf{y}}_t^{\,\top} \widetilde{\mathbf{C}} \widetilde{\mathbf{y}}_t \leq 2 \left(1 + \frac{1}{c}\right) \cdot (1 + \mathsf{SU}_b(t)^2) \cdot \mathsf{tr}(\widetilde{\mathbf{C}}) \cdot \ln n$$

almost surely for every realization of the random matrix $\widetilde{\mathbf{C}}$, and with probability at least $\left(1 - \frac{2}{n}\right)$ over the randomness in $\widetilde{\mathbf{y}}_t$.

Applying Lemma 28, we get

$$\mathsf{CN}_b^2(t) \leq 2\left(1+\frac{1}{c}\right)\cdot \mathsf{tr}(\widetilde{\mathbf{C}})\cdot \ln n \tag{54}$$

almost surely for every realization of the random matrix $\widetilde{\mathbf{C}}$, and with probability at least $\left(1-\frac{2}{n}\right)$ over the randomness in $\widetilde{\mathbf{y}}_t$. The next lemma, which is taken from Bartlett et al. (2020), provides a high-probability upper bound on the quantity $\mathsf{tr}(\widetilde{\mathbf{C}})$.

Lemma 29 *(From Lemma 11 in Bartlett et al. (2020)) There exist universal constants $(b_2, c_5, c_{10} \geq 1)$ such that whenever $0 \leq k \leq n/c_5$ and $r_k(\mathbf{\Sigma}_{-t}) \geq b_2 n$, we have*

$$\mathsf{tr}(\widetilde{\mathbf{C}}) \leq c_{10}\cdot\left(\frac{l}{n}+n\cdot\frac{\sum_{j>l}\widetilde{\lambda}_j^2}{\left(\sum_{j>k}\widetilde{\lambda}_j\right)^2}\right)$$

for any choice of $l \leq k$, with probability at least $(1-6e^{-\frac{n}{c_5}})$ over the randomness in $\widetilde{\mathbf{C}}$.

Substituting the upper bound from Lemmas 29 and into Equation (54), and taking the square root on both sides, we have

$$\mathsf{CN}_b(t) \leq \sqrt{2\left(1+\frac{1}{c}\right)\cdot c_{10}\cdot\left(\frac{l}{n}+n\cdot\frac{\sum_{j>l}\widetilde{\lambda}_j^2}{\left(\sum_{j>k}\widetilde{\lambda}_j\right)^2}\right)\cdot(1+\mathsf{SU}_b(t)^2)\cdot\ln n}.$$

with probability at least $\left(1-\frac{2}{n}-6e^{-\frac{n}{c_2}}\right)$ over the training data. Taking $c_7 = \sqrt{2\left(1+\frac{1}{c}\right)c_{10}}$, the upper bound on $\mathsf{CN}_b(t)$ in Equation (35a) follows. Noting that $\left(1-\frac{2}{n}-6e^{-\frac{n}{c_2}}\right) \geq \left(1-\frac{3}{n}\right)$ for large enough n, this completes the proof of the upper bound.

D.4.3 Lower bound on $\mathsf{CN}_b(t)$

Now we move on to the proof for the lower bound on contamination for interpolation of binary labels (Equation (35b)). Using Equation (53a) from Lemma 27, we get

$$\begin{aligned}\mathsf{CN}_b^2(t) &= \mathbf{y}_t^\top \mathbf{C}\mathbf{y}_t \\ &\geq \mu_n(\mathbf{C})\,\|\mathbf{y}_t\|_2^2 = n\mu_n(\mathbf{C}).\end{aligned}$$

The next lemma lower bounds the minimum eigenvalue of $\mathbf{C}$ and is proved in Appendix F.5.

Lemma 30 *Let $k \geq 0$ and $r_k\left(\mathbf{\Sigma}_{-t}^2\right) \geq b_4 n$. Then, we have*

$$\mu_n(\mathbf{C}) \geq \frac{r_k\left(\mathbf{\Sigma}_{-t}^2\right)\cdot\widetilde{\lambda}_{k+1}^2}{c_{11}\cdot c^2\cdot\left(\sum_{j=1}^{d}\lambda_j+\lambda_1 n\right)^2}$$

with probability at least $(1-e^{-\frac{n}{c}}-e^{-\frac{n}{c_{11}}})$.= over the randomness in $\mathbf{C}$. Here, (b_4, c, c_{11}) are universal positive constants.

A direct substitution of the above gives us

$$\mathsf{CN}_b(t) \geq \sqrt{n} \cdot \frac{\sqrt{r_k\left(\mathbf{\Sigma}_{-t}^2\right) \cdot \widetilde{\lambda}_{k+1}^2}}{c \cdot \sqrt{c_{11}} \cdot \left(\sum_{j=1}^{d} \lambda_j + \lambda_1 n\right)}$$

with probability at least $(1 - e^{-\frac{n}{c}} - e^{-\frac{n}{c_{11}}})$ over the training data. Taking $c_9 = c\sqrt{c_{11}}$ and c_8 such that $\frac{1}{c_8} = \min(\frac{1}{c}, \frac{1}{c_{11}})$ holds, the lower bound in Equation (35b) follows. This completes the characterization of the contamination factor when we interpolate binary labels.

D.4.4 Interpolation of real output

For completeness, we also provide the proof of Theorem 23 for the simpler case of interpolation of real output. We start with a useful set of expressions for the contamination factor in the following lemma. The proof of this lemma is contained in Appendix F.3.

Lemma 31 *The contamination of the minimum-ℓ_2-norm interpolator of binary labels, denoted by $\widehat{\boldsymbol{\alpha}}_{2,\mathsf{real}}$, can be written in the following two forms:*

$$\mathsf{CN}_r(t) = \sqrt{\mathbf{z}_t^\top \mathbf{C} \mathbf{z}_t}, \tag{55a}$$

$$= |1 - \mathsf{SU}_r(t)| \sqrt{\mathbf{z}_t^\top \widetilde{\mathbf{C}} \mathbf{z}_t}, \tag{55b}$$

where we denote

$$\mathbf{C} = \mathbf{A}^{-1} \left(\sum_{j=1, j\neq t}^{d} \lambda_j^2 \mathbf{z}_j \mathbf{z}_j^\top \right) \mathbf{A}^{-1} \text{ , and}$$

$$\widetilde{\mathbf{C}} = \mathbf{A}_{-t}^{-1} \left(\sum_{j=1, j\neq t}^{d} \lambda_j^2 \mathbf{z}_j \mathbf{z}_j^\top \right) \mathbf{A}_{-t}^{-1}.$$

We will use the form in Equation (55b) to prove an upper bound on contamination and the form in Equation (55a) for the lower bound.

D.4.5 Upper bound on $\mathsf{CN}_r(t)$

We start with the proof for the upper bound on contamination for interpolation of real output (Equation (36a). From Equation (55a) in Lemma 31, we get

$$\mathsf{CN}_r^2(t) = (1 - \mathsf{SU}_r(t))^2 \mathbf{z}_t^\top \widetilde{\mathbf{C}} \mathbf{z}_t. \tag{56}$$

From Equation (75) in Appendix F.4 (proof of Lemma 28), we can upper bound the quadratic form $\mathbf{z}_t^\top \widetilde{\mathbf{C}} \mathbf{z}_t$ as

$$\mathbf{z}_t^\top \widetilde{\mathbf{C}} \mathbf{z}_t \leq 7\mathsf{tr}(\widetilde{\mathbf{C}}) \ln n$$

with probability at least $\left(1 - \frac{1}{n}\right)$ over the randomness in $\mathbf{z}_t$. Then, substituting the upper bound on $\mathsf{tr}(\widetilde{\mathbf{C}})$ from Lemma 29 directly gives us the expression for the upper bound on $\mathsf{CN}_r(t)$. Noting again that $\left(1 - \frac{1}{n} - 6e^{-\frac{n}{c_2}}\right) \geq \left(1 - \frac{2}{n}\right)$ for large enough n, this completes the proof for the upper bound.

D.4.6 Lower bound

We conclude this section by proving the lower bound on contamination for interpolation of real output (Equation (36b)). We directly apply Equation (55a) (from Lemma 31) to get

$$\begin{aligned}\mathsf{CN}_r^2(t) &= \mathbf{z}_t^\top \mathbf{C} \mathbf{z}_t \\ &\geq \mu_n(\mathbf{C}) \left\|\mathbf{z}_t\right\|_2^2 \\ &\overset{\text{(i)}}{\geq} n(1-\delta)\mu_n(\mathbf{C})\end{aligned}$$

with probability at least $(1 - e^{-n\delta^2})$ over the randomness in $\mathbf{z}_t$ for any $\delta \in (0, 1)$. Here, inequality (i) follows from the lower bound in Lemma 26. Finally, substituting the lower bound for $\mu_n(\mathbf{C})$ from Lemma 30 gives us the desired expression for the lower bound on $\mathsf{CN}_r(t)$. Note that by the union bound, this expression will hold with probability at least $(1 - e^{-n\delta^2} - e^{-\frac{n}{c}} - e^{-\frac{n}{c_{11}}}) = (1 - 2e^{-\frac{n}{c_8}} - e^{-n\delta^2})$ over the randomness in the training data. This completes the proof of Theorem 23. ■

Appendix E. Implications for bi-level covariance: Proof of Theorem 13

In this section, we follow the *path to analysis* described in Section 5.2 and prove Theorem 13 for the bi-level ensemble (Definition 5) in the following series of steps:

1. We substitute the spectrum of the bi-level ensemble into Theorems 22 and 23 to get asymptotic expressions for survival and contamination.
2. We substitute these expressions into the expressions for regression and classification test loss (Proposition 17) to characterize the regimes for good generalization of classification and regression.

For convenience of notation, we consider $t = 1$. (Note, however, that the analysis holds for any $1 \leq t \leq s$ since the first s eigenvalues of $\mathbf{\Sigma}$ are equal.) Further, to emphasize that the survival and contamination quantities depend on n, in this section we refer to them as $\mathsf{SU}_b(1; n), \mathsf{CN}_b(1; n), \mathsf{SU}_r(1; n),$ and $\mathsf{CN}_r(1; n)$ for interpolators of binary and real output respectively.

First, we characterize some useful quantities for the bi-level ensemble. Recall that the bi-level ensemble is parameterized by $p > 1$, $0 < q \leq (p - r)$ and $0 < r \leq 1$. We first compute the effective ranks $r_k(\mathbf{\Sigma})$ and $r_k(\mathbf{\Sigma}_{-t})$ for two choices of k. First, we have

$$r_s(\mathbf{\Sigma}) = \frac{1}{\frac{(1-a)d}{d-s}} . \frac{(1-a)d}{d-s} . (d-s) = d - s.$$

Substituting $d = n^p$ and $s = n^r$, we have, for sufficiently large n,

$$r_s(\mathbf{\Sigma}) \asymp n^p \gg n. \tag{57}$$

Similarly because $1 \leq t \leq s$, we have, for sufficiently large n,

$$r_s(\mathbf{\Sigma}_{-t}) = d - s - 1 \asymp n^p \gg n. \tag{58}$$

Moreover, we get

$$r_0(\mathbf{\Sigma}) = \frac{1}{\frac{ad}{s}} \cdot d = \frac{s}{a} = n^{q+r} \gg n \text{ iff } (q+r) > 1. \tag{59}$$

and by a similar argument, provided that $r > 0$, we can show that (for large enough n),

$$r_0(\mathbf{\Sigma}_{-t}) = \frac{1}{\frac{ad}{s}} \cdot \left(d - \frac{ad}{s}\right) = \frac{s}{a} - 1 = n^{q+r} - 1 \gg n \text{ iff } (q+r) > 1. \tag{60}$$

We will apply Equations (57) and (58) for bounding survival in general, as well as contamination when we have $q \leq (1-r)$, and Equations (59) and (60) for bounding contamination when we have $q > (1-r)$. Now, we state and prove our matching upper and lower bounds for survival for the bi-level ensemble.

Lemma 32 (Survival for interpolation of binary labels) *There exist universal positive constants (L_1, U_1, L_2, U_2) such that for sufficiently large n, we have*

$$\mathsf{SU}_b^L(n) \leq \mathsf{SU}_b(1; n) \leq \mathsf{SU}_b^U(n),$$

with probability at least $(1 - 10e^{-\sqrt{n}})$ over the training data $\{X_i, Y_i\}_{i=1}^n$, where we denote

$$\mathsf{SU}_b^L(n) := \begin{cases} \sqrt{\frac{2}{\pi}}(1-2\nu^*)\left(1 + L_1 n^{q-(1-r)}\right)^{-1}, & q < (1-r) \\ \sqrt{\frac{2}{\pi}}(1-2\nu^*) \cdot L_2 n^{(1-r)-q}, & q > (1-r) \end{cases}, \tag{61a}$$

$$\mathsf{SU}_b^U(n) := \begin{cases} \sqrt{\frac{2}{\pi}}(1-2\nu^*)\left(1 + U_1 n^{q-(1-r)}\right)^{-1}, & q < (1-r) \\ \sqrt{\frac{2}{\pi}}(1-2\nu^*) \cdot U_2 n^{(1-r)-q}, & q > (1-r) \end{cases}. \tag{61b}$$

Proof Note that Equations (57) and (58) imply that the conditions $r_s(\mathbf{\Sigma}) \geq bn$ and $r_s(\mathbf{\Sigma}_{-t}) \geq b_2 n$ are clearly satisfied for large enough n. Thus, we can apply Equation (33a) of Theorem 22 setting $k = s$ to get

$$\mathsf{SU}_b(1; n) \geq \sqrt{\frac{2}{\pi}}(1-2\nu^*) \frac{\lambda_1\left(\frac{(n-s)}{c\widetilde{\lambda}_{s+1} r_s(\mathbf{\Sigma}_{-1})} - \frac{c_3 n^{3/4}}{\lambda_{s+1} r_s(\mathbf{\Sigma})}\right)}{1 + \lambda_1\left(\frac{cn}{\widetilde{\lambda}_{s+1} r_s(\mathbf{\Sigma}_{-1})} + \frac{c_4 n^{3/4}}{\lambda_{s+1} r_s(\mathbf{\Sigma})}\right)}$$

with probability at least $(1 - 5e^{-\sqrt{n}})$ over the training data. Substituting $s = n^r$ and $a = n^{-q}$, note that

$$\frac{\lambda_{s+1} r_s(\mathbf{\Sigma})}{\lambda_1} = \frac{\widetilde{\lambda}_{s+1} r_s(\mathbf{\Sigma}_{-1})}{\lambda_1} \asymp \frac{\frac{(1-\gamma)d}{d-s} n^p}{\frac{\gamma d}{s}} \asymp \frac{n^{p+r}}{n^{p-q}} \asymp n^{q+r}.$$

Substituting this above yields

$$\begin{aligned} \mathsf{SU}_b(1; n) &\geq \sqrt{\frac{2}{\pi}}(1-2\nu^*)\left(\frac{\frac{(n-n^r)}{cn^{q+r}} - \frac{c_3 n^{3/4}}{n^{q+r}}}{1 + \frac{cn}{n^{q+r}} + \frac{c_4 n^{3/4}}{n^{q+r}}}\right) \\ &= \sqrt{\frac{2}{\pi}}(1-2\nu^*)\left(\frac{\frac{1}{c} \cdot (n^{(1-r)-q} - n^{-q}) - c_3 \cdot n^{(3/4-r)-q}}{1 + cn^{(1-r)-q} + c_4 \cdot n^{(3/4-r)-q}}\right). \end{aligned}$$

Thus, there are two cases:

1. $0 < q \leq (1-r)$, in which case the terms corresponding to $n^{q-(1-r)}$ dominate, and there exists universal constant L_1 such that
$$\mathsf{SU}_b(1;n) \geq \sqrt{\frac{2}{\pi}}(1-2\nu^*)\left(1+L_1 n^{q-(1-r)}\right)^{-1}.$$
2. $q > (1-r)$, in which case the numerator goes to 0 but the denominator goes to 1 as $n \to \infty$, and so there exists universal constant L_2 such that
$$\mathsf{SU}_b(1;n) \geq \sqrt{\frac{2}{\pi}}(1-2\nu^*) \cdot L_2 n^{(1-r)-q}.$$

This completes the proof of the lower bound. An almost identical argument gives the proof of the upper bound, so we omit it here. ■

Observe that for $q > (1-r)$, the true signal does not survive at all, i.e. $\mathsf{SU}_r(1;n) \to 0$ as $n \to \infty$. Interestingly, for $q \leq (1-r)$, there is also non-trivial attenuation of signal when binary labels are interpolated, i.e. $\mathsf{SU}_r(1;n) \to \sqrt{\frac{2}{\pi}} \cdot (1-2\nu^*) < 1$ as $n \to \infty$. At a high level, this is a consequence of effective misspecification induced by the sign operator on real output. As mentioned in the discussion in Section 5.2, this is also spiritually related to the attenuation factor of signal that has been traditionally been observed as a result of 1-bit quantization applied to a matched filter (Turin, 1976; Chang, 1982).

As we will see in the following lemma, the corresponding case leads to zero attenuation of signal when real output is interpolated., i.e. $\mathsf{SU}_r(1;n) \to 1$.

Lemma 33 (Survival for interpolation of real output) *There exist universal positive constants $(L_1, U_1, L_2, U_2, \overline{L_1}, \overline{U_1}, \overline{L_2}, \overline{U_2})$ such that for sufficiently large n, we have*
$$\mathsf{SU}_r^L(n) \leq \mathsf{SU}_r(1;n) \leq \mathsf{SU}_r^U(n),$$
with probability at least $(1-8e^{-\sqrt{n}})$ over the randomness in the training data $\{X_i, Y_i\}_{i=1}^n$, where we denote
$$\mathsf{SU}_r^L(n) := \begin{cases} \left(1+L_1 n^{q-(1-r)}\right)^{-1}, & q < (1-r) \\ L_2 n^{(1-r)-q}, & q > (1-r) \end{cases}, \tag{62a}$$
$$\mathsf{SU}_r^U(n) := \begin{cases} \left(1+U_1 n^{q-(1-r)}\right)^{-1}, & q < (1-r) \\ U_2 n^{(1-r)-q}, & q > (1-r) \end{cases}. \tag{62b}$$
Equivalently, we can write
$$\overline{\mathsf{SU}_\mathsf{r}}^L(n) \leq 1 - \mathsf{SU}_r(1;n) \leq \overline{\mathsf{SU}_\mathsf{r}}^U(n),$$
where we denote
$$\overline{\mathsf{SU}_\mathsf{r}}^L(n) := \begin{cases} \overline{L_1} n^{q-(1-r)}, & q < (1-r) \\ \left(1+\overline{L_2} n^{(1-r)-q}\right)^{-1}, & q > (1-r) \end{cases}, \tag{63a}$$
$$\overline{\mathsf{SU}_\mathsf{r}}^U(n) := \begin{cases} \overline{U_1} n^{q-(1-r)}, & q < (1-r) \\ \left(1+\overline{U_2} n^{(1-r)-q}\right)^{-1}, & q > (1-r) \end{cases} \tag{63b}$$

Proof The proof follows by substituting the spectrum of the bi-level covariance model into the upper and lower bounds of survival from Equations (34b) and (34a). This is essentially an identical argument to the proof of Lemma 32, and so we omit it here. ■

Observe that for the case of interpolation of real output, we have additionally computed bounds on the quantity $(1-\mathsf{SU}_r(1;n))$, which will subsequently be useful for the computation of bounds on contamination. We have not stated this here to avoid complicating the proof, but it is interesting to note that if the real-valued output had a non-zero level of independent additive zero-mean Gaussian noise, then this would not matter for the scaling of the survival results asymptotically — this is a consequence of the range of parameter choices that we have chosen for our bi-level ensemble. Such label noise would effectively be completely absorbed by the excess features.

We now state an upper bound on contamination for the bi-level ensemble.

Lemma 34 (Contamination for interpolation of binary labels) *There are universal positive constants $(U_3, U_4$ and $U_5)$ such that for large enough n, we have $\mathsf{CN}_b(1;n) \leq \mathsf{CN}_b^U(n)$ with probability at least $\left(1-\frac{4}{n}\right)$ over the randomness in the training data $\{X_i, Y_i\}_{i=1}^n$, where we denote*

$$\mathsf{CN}_b^U(n) = \begin{cases} U_3 n^{\frac{-\min\{(p-1),(1-r)\}}{2}} \cdot \sqrt{\ln n} \text{ if } q < (1-r) \\ U_4 n^{\frac{-\min\{(p-1),(2q+r-1)\}}{2}} \cdot \sqrt{\ln n} \text{ if } q > (1-r) \end{cases} \tag{64}$$

Proof We start by proving the statement for the case $q \leq (1-r)$. From Equations (57) and (58), we showed that for large enough n, we have $r_s(\boldsymbol{\Sigma}_{-1}) \asymp n^p \gg n$. Substituting $k = l = s$ in Equation (35a) from Theorem 23, we have

$$\mathsf{CN}_b(1;n) \leq c_7 \cdot \sqrt{\left(\frac{s}{n} + n \cdot \frac{\sum_{j>s} \widetilde{\lambda}_j^2}{\left(\sum_{j>s} \widetilde{\lambda}_j\right)^2}\right) \cdot \ln n \cdot (1 + \mathsf{SU}_b(1;n)^2)} \tag{65}$$

almost surely for every realization of SU with probability at least $\left(1-\frac{3}{n}\right)$ over the training data. We first evaluate the term

$$T_1 := \frac{s}{n} + n \cdot \frac{\sum_{j>s} \widetilde{\lambda}_j^2}{\left(\sum_{j>s} \widetilde{\lambda}_j\right)^2}.$$

First, note that

$$\sum_{j>s} \widetilde{\lambda}_j^2 = (d-s-1)\left(\frac{(1-\gamma)d}{d-s}\right)^2 \asymp d = n^p \text{ and}$$

$$\left(\sum_{j>s} \widetilde{\lambda}_j\right)^2 = \left((d-s-1)\frac{(1-\gamma)d}{d-s}\right)^2 \asymp n^{2p}.$$

Using this, we obtain

$$T_1 \asymp n^{(r-1)} + n^{(1-p)} \asymp n^{-\min\{(p-1),(1-r)\}}. \tag{66}$$

Now, from Equation (61b), we get (for large enough n)

$$\mathsf{SU}_b(1;n) \leq \mathbb{1}_{q\leq(1-r)}\sqrt{\frac{2}{\pi}}\left(1+U_1 n^{q-(1-r)}\right)^{-1} + \mathbb{1}_{q>(1-r)}U_2 n^{(1-r)-q} \leq \max\left\{U_2, \sqrt{\frac{2}{\pi}}\right\} \tag{67}$$

with probability at least $(1-4e^{-p_1 n})$ over the training data. Substituting Equations (66) and (67) in Equation (65), we have

$$\mathsf{CN}_b(1;n) \leq U_3 n^{-\frac{\min\{(p-1),(1-r)\}}{2}} \cdot \sqrt{\ln n}$$

with probability at least $\left(1-\frac{4}{n}\right)$ for appropriately defined positive constant U_3. This completes the proof for the first case.
Now, we move on to the second case, i.e. $q > (1-r)$. From Equations (59) and (60), we saw that in this case, we have $r_0(\mathbf{\Sigma}_{-1}) \asymp n^{q+r} \gg n$. Substituting $k = l = 0$ in Equation (35a) from Theorem 23, we have

$$\mathsf{CN}_b(1;n) \leq c_7 \cdot \sqrt{\left(n \cdot \frac{\sum_{j>0}\widetilde{\lambda}_j^2}{\left(\sum_{j>0}\widetilde{\lambda}_j\right)^2}\right) \cdot \ln n \cdot (1+\mathsf{SU}_b(1;n)^2)}$$

with probability at least $\left(1-\frac{3}{n}\right)$ over the training data. As before, we evaluate the term

$$T_1 := n \cdot \frac{\sum_{j>0}\widetilde{\lambda}_j^2}{(\sum_{j>0}\widetilde{\lambda}_j)^2}$$

By a calculation very similar to the one in Appendix C.2, we get

$$\sum_{j>0}\widetilde{\lambda}_j^2 = (s-1)\cdot\frac{a^2d^2}{s^2} + (d-s-1)\cdot\frac{(1-a)^2d^2}{(d-s)^2} \asymp n^{2p+2q-r} + n^p.$$

Moreover, we get $(\sum_{j>0}\widetilde{\lambda}_j)^2 = (d-\frac{ad}{s})^2 = (n^p - n^{p-(r+q)})^2 \asymp n^{2p}$ since $(q+r) > 0$. Therefore, we get

$$T_1 \asymp n^{(1-p)} + n^{(1+2q-r)} \asymp n^{-\min\{(p-1),(2q+r-1)\}}.$$

The other steps proceed as for the first case, and substituting this expression for the term T_1 completes the proof for the second case. ■

For some parameterizations of the bi-level ensemble, we can get a slightly more sophisticated upper bound on contamination when the labels interpolated are real, as detailed in the following lemma.

Lemma 35 (Contamination for interpolation of real output) *For universal positive constants (U_3, U_4, U_5) and large enough n, we have $\mathsf{CN}_r(1;n) \leq \mathsf{CN}_r^U(n)$ with probability at least $\left(1 - \frac{3}{n}\right)$ over the randomness in the training data $\{X_i, Y_i\}_{i=1}^n$, where we denote*

$$\mathsf{CN}_r^U(n) = \begin{cases} U_3 n^{q-(1-r)-\frac{\min\{(p-1),(1-r)\}}{2}} \cdot \sqrt{\ln n}, & q < (1-r), \\ U_4 n^{-\frac{\min\{(p-1),(2q+r-1)\}}{2}} \cdot \sqrt{\ln n}, & q > (1-r) \end{cases}. \tag{68}$$

Proof We follow an identical approach as in the proof of Lemma 34 to bound the term T_1. Substituting this along with the upper bound on the quantity $(1 - \mathsf{SU}_r(1;n))$ from Equation (63b) (Lemma 33) in Equation (36a), and using the fact that $\mathsf{SU}_r(1;n) \leq 1$, Equation (68) follows for appropriately defined positive constants (U_3, U_4). This completes the proof. ■

Finally, we state and prove our lower bounds on contamination together for interpolation of binary labels as well as real output.

Lemma 36 (Lower bounds on contamination) *There are universal positive constants (L_3, L_4, p_2) such that for large enough n, we have $\mathsf{CN}_b(1;n), \mathsf{CN}_r(1;n) \geq \mathsf{CN}^L(n)$ with probability at least $(1 - 2e^{-p_2 n})$ over the randomness in the training data $\{X_i, Y_i\}_{i=1}^n$, where we define*

$$\mathsf{CN}^L(n) := \begin{cases} L_3 n^{q-(1-r)-\frac{p-1}{2}}, & q < (1-r) \\ L_4 n^{-\frac{(p-1)}{2}}, & q > (1-r) \end{cases}. \tag{69}$$

Proof Using Equation (58) we have, for large enough n, $r_s\left(\mathbf{\Sigma}_{-1}^2\right) \asymp n^p \gg n$. Taking $k = s$ in Equation (35b) from Theorem 23, for universal constants c_8, c_9, with probability at least $(1 - 2e^{-\frac{n}{c_8}})$, we have

$$\begin{aligned} \mathsf{CN}_b(1;n) &\geq \sqrt{n}. \frac{\sqrt{r_s\left(\mathbf{\Sigma}_{-1}^2\right) \widetilde{\lambda}_{s+1}^2}}{c_9 \left(\sum_{j=1}^{d} \lambda_j + \lambda_1 n \right)} \\ &\asymp n^{\frac{1}{2}} \cdot \frac{\sqrt{n^p \left(\frac{(1-\gamma)d}{d-s} \right)^2}}{d + n^{\frac{\gamma d}{s}}} \\ &\asymp \frac{n^{-\frac{(p-1)}{2}}}{1 + n^{(1-r)-q}}, \\ &\asymp \begin{cases} n^{q-(1-r)-\frac{(p-1)}{2}}, & q < (1-r) \\ n^{-\frac{(p-1)}{2}}, & q > (1-r) \end{cases}. \end{aligned}$$

Thus Equation (64) follows by choosing appropriate constants p_2, L_3 and L_4, completing the proof.

■

Comparing the upper bound (Equation (68)) and lower bound (Equation (69)) for the case of interpolating real output, we observe that these bounds would be matching up to constant factors *iff* $(p-1) \leq (1-r)$. In addition to the above condition, the upper bound for interpolation of binary labels (Equation (65)) will match the lower bound *iff* $q > (1-r)$.

Finally, we compute bounds on the ratio of survival to contamination, $\mathsf{SU}_b(1;n)/\mathsf{CN}_b(1;n)$, for the interpolation of binary labels. A directly substitution of the upper and lower bounds for $\mathsf{SU}_b(1;n)$ and $\mathsf{CN}_b(1;n)$ from Equations (61a), (61b) in Lemma 32, Equations (64) in Lemma 34 and Equation (69) in Lemma 36, gives us (for large enough n)

$$\mathsf{SNR}^L(n) \leq \frac{\mathsf{SU}_b(1;n)}{\mathsf{CN}_b(1;n)} \leq \mathsf{SNR}^U(n), \tag{70}$$

with probability at least $\left(1 - \frac{16}{n}\right)$ over the training data, where we denote

$$\mathsf{SNR}^L(n) := \begin{cases} L_5 \cdot n^{\frac{\min\{(p-1),(1-r)\}}{2}} \cdot (\ln n)^{-\frac{1}{2}}, & 0 < q < (1-r) \\ L_6 \cdot n^{\frac{\min\{(p-1),(2q+r-1)\}}{2} + (1-r) - q} \cdot (\ln n)^{-\frac{1}{2}}, & q > (1-r) \end{cases}. \tag{71a}$$

$$\mathsf{SNR}^U(n) = U_5 \cdot n^{\frac{p-1}{2} + (1-r) - q}. \tag{71b}$$

E.1 Proof of Theorem 13

We are now ready to complete the proof of Theorem 13. First we compute a lower bound on regression test loss. From Equations (17), (63a) and (69), we have (for large enough n)

$$\begin{aligned} \mathcal{R}(\widehat{\boldsymbol{\alpha}}_{2,\mathsf{real}}; n) &= (1 - \mathsf{SU}_r(1;n))^2 + (\mathsf{CN}_r(1;n))^2 \\ &\geq (\overline{\mathsf{SU}_\mathsf{r}}^L(n))^2 + (\mathsf{CN}_r^L(n))^2 \\ &= \begin{cases} \overline{L_1}^2 n^{2(q-(1-r))} + L_3^2 n^{-2(1-r)-(p-1)+2q}, & q < (1-r) \\ \left(1 + \overline{L_2} n^{(1-r)-q}\right)^{-2} + L_4^2 n^{-(p-1)}, & q > (1-r) \end{cases} \end{aligned}$$

with probability at least $(1 - 2e^{-\sqrt{n}} - 2e^{-p_2 n})$. Thus, we have

$$\liminf_{n\to\infty} \mathcal{R}(\widehat{\boldsymbol{\alpha}}_{2,\mathsf{real}}; n) \geq \begin{cases} 0, & q < (1-r) \\ 1, & q > (1-r) \end{cases}$$

with probability equal to 1. Next, we compute an upper bound on regression test loss. From Equations (17), (63b) and (68), we have (for large enough n)

$$\begin{aligned} \mathcal{R}(\widehat{\boldsymbol{\alpha}}_{2,\mathsf{real}}; n) &\leq (\overline{\mathsf{SU}_\mathsf{r}}^U(n))^2 + (\mathsf{CN}_r^U(n))^2 \\ &= \begin{cases} \overline{U_1}^2 n^{2(q-(1-r))} + U_3^2 n^{-2(1-r)-\min\{(p-1),(1-r)\}+2q} \ln n, & q < (1-r) \\ \left(1 + \overline{U_2} n^{(1-r)-q}\right)^{-2} + U_4^2 n^{-(p-1)} \ln n, & q > (1-r) \end{cases} \end{aligned}$$

with probability at least $\left(1 - 2e^{-\sqrt{n}} - \frac{3}{n}\right)$. Thus, we have

$$\limsup_n \mathcal{R}(\widehat{\boldsymbol{\alpha}}_{2,\mathsf{real}}; n) \leq \begin{cases} 0, & q < (1-r) \\ 1, & q > (1-r) \end{cases}$$

with probability equal to 1. By the sandwich theorem, we get

$$\lim_{n\to\infty} \mathcal{R}(\widehat{\boldsymbol{\alpha}}_{2,\mathsf{real}}; n) = \begin{cases} 0, & q < (1-r) \\ 1, & q > (1-r) \end{cases}$$

with probability 1, completing our characterization of regression.
We now move on to our final characterization of classification test loss, starting with the upper bound. By Proposition 17, we have

$$\mathcal{C}(\widehat{\boldsymbol{\alpha}}_{2,\mathsf{binary}}; n) = \frac{1}{2} - \frac{1}{\pi}\mathsf{tan}^{-1}\left(\frac{\mathsf{SU}_b(1;n)}{\mathsf{CN}_b(1;n)}\right).$$

From Equation (70), we get

$$\frac{1}{2} - \frac{1}{\pi}\mathsf{tan}^{-1}\left(\mathsf{SNR}^U(n)\right) \leq \mathcal{C}(\widehat{\boldsymbol{\alpha}}_{2,\mathsf{binary}}; n) \leq \frac{1}{2} - \frac{1}{\pi}\mathsf{tan}^{-1}\left(\mathsf{SNR}^L(n)\right).$$

Taking the limit as $n \to \infty$ in Equation (71a), we have

$$\liminf_{n\to\infty} \mathsf{SNR}^L(n) = \begin{cases} \infty, & q < \frac{\min\{(p-1),(2q+r-1)\}}{2} + (1-r) \\ 0, & q > \frac{\min\{(p-1),(2q+r-1)\}}{2} + (1-r) \end{cases}.$$

with probability 1. Thus, we have

$$\limsup_n \mathcal{C}(\widehat{\boldsymbol{\alpha}}_{2,\mathsf{binary}}; n) \leq \begin{cases} 0, & q < \frac{\min\{(p-1),(2q+r-1)\}}{2} + (1-r) \\ \frac{1}{2}, & q > \frac{\min\{(p-1),(2q+r-1)\}}{2} + (1-r) \end{cases}.$$

with probability 1. To simplify further, consider the case for which $(2q + r - 1) < (p - 1)$. Then, the condition becomes $q < q + \frac{(r-1)}{2} + (1-r) = \frac{(1-r)}{2} \implies \frac{(1-r)}{2} > 0$, which is always true under the bi-level ensemble (as $r < 1$). Thus, we can effectively ignore this argument, and simply write

$$\limsup_n \mathcal{C}(\widehat{\boldsymbol{\alpha}}_{2,\mathsf{binary}}; n) \leq \begin{cases} 0, & q < \frac{(p-1)}{2} + (1-r) \\ \frac{1}{2}, & q > \frac{(p-1)}{2} + (1-r) \end{cases}.$$

On the other hand, we can also compute the limiting upper bound on SNR:

$$\limsup_n \mathsf{SNR}^U(n) = \begin{cases} \infty, & 0 < q < \frac{(p-1)}{2} + (1-r) \\ 0, & q > \frac{(p-1)}{2} + (1-r). \end{cases}$$

and so the classification test loss is *lower bounded* by:

$$\liminf_n \mathcal{C}(\widehat{\boldsymbol{\alpha}}_{2,\mathsf{binary}}; n) \geq \begin{cases} 0, & 0 < q < \frac{(p-1)}{2} + (1-r) \\ \frac{1}{2}, & q > \frac{(p-1)}{2} + (1-r). \end{cases}$$

Putting these together, we get

$$\lim_{n\to\infty} \mathcal{C}(\widehat{\boldsymbol{\alpha}}_{2,\mathsf{binary}}; n) = \begin{cases} 0, & 0 < q < \frac{(p-1)}{2} + (1-r) \\ \frac{1}{2}, & q > \frac{(p-1)}{2} + (1-r). \end{cases}$$

This completes the proof. ■

Appendix F. Technical lemmas

F.1 Proof of Lemma 20

In this subsection, we prove Lemma 20, i.e. concentration on the operator norm of the random matrix $\mathbf{E} := \mathbf{A} - ||\boldsymbol{\lambda}||_1 \mathbf{I}_n$. Recall that $\mathbf{A}$ is the random Gram matrix as defined in Appendix B. It is easy to verify that $\mathbb{E}[\mathbf{A}] = ||\boldsymbol{\lambda}||_1 \mathbf{I}_n$. We start by recalling the following lemma by Laurent and Massart (Laurent and Massart, 2000).

Lemma 37 ***(Laurent and Massart, 2000)*** *For any $t > 0$, and any $\mathbf{u} \in \mathbb{R}^n$, we have*

$$\Pr\left[\mathbf{u}^\top \mathbf{E}\mathbf{u} > \sqrt{2||\boldsymbol{\lambda}||_2 \cdot t} + 2||\boldsymbol{\lambda}||_\infty \cdot t\right] \leq e^{-t}$$
$$\Pr\left[\mathbf{u}^\top \mathbf{E}\mathbf{u} < -\sqrt{2||\boldsymbol{\lambda}||_2 \cdot t}\right] \leq e^{-t},$$

where the probability is taken over the randomness in the matrix $\mathbf{E}$.

We use this lemma together with a discretization and covering argument. Let $\mathcal{U} := \{\mathbf{u}_1, \ldots, \mathbf{u}_N\}$ be an ϵ-net for the unit sphere $\mathcal{S}^{n-1}$ in $\mathbb{R}^n$, i.e., we have $\min_{i\in\{1,\ldots,N\}} ||\mathbf{u} - \mathbf{u}_i||_2 \leq \epsilon$ for all $\mathbf{u} \in \mathcal{S}^{n-1}$. It is easy to show (for e.g. according to the covering arguments provided in Wainwright (2019, Chapter 4)) that we can pick $\mathcal{U}$ to be an ϵ-net such that $N \leq \left(1 + \frac{2}{\epsilon}\right)^n$.

We set $t := \tau + n \ln\left(1 + \frac{2}{\epsilon}\right)$ for some $\tau > 0$. Then, by a union bound over the set $\mathcal{U}$, we have

$$\max_{i\in[N]} \mathbf{u}_i^\top \mathbf{E}\mathbf{u}_i \leq \sqrt{2||\boldsymbol{\lambda}||_2 \cdot t} + 2||\boldsymbol{\lambda}||_\infty \cdot t \text{ and} \tag{72a}$$
$$\min_{i\in[N]} \mathbf{u}_i^\top \mathbf{E}\mathbf{u}_i \geq -\sqrt{2||\boldsymbol{\lambda}||_2 \cdot t} \tag{72b}$$

with probability at least $(1 - 2e^{-\tau})$ over the randomness in the matrix $\mathbf{E}$. It now remains to remove the discretization in both directions. Let $\widehat{\mathbf{u}} := \arg\max_{\mathbf{u}\in\mathcal{S}^{n-1}} \mathbf{u}^\top \mathbf{A}\mathbf{u} = \arg\max_{\mathbf{u}\in\mathcal{S}^{n-1}} ||\mathbf{A}^{1/2}\mathbf{u}||_2$, and let $i_0 := \arg\min_{i\in\{1,\ldots,N\}} ||\widehat{\mathbf{u}} - \mathbf{u}_i||_2$ denote the nearest neighbor of $\widehat{\mathbf{u}}$. Then, we have

$$\begin{aligned} ||\mathbf{A}^{1/2}\widehat{\mathbf{u}}||_2 &= ||\mathbf{A}^{1/2}\mathbf{u}_{i_0} + \mathbf{A}^{1/2}(\widehat{\mathbf{u}} - \mathbf{u}_{i_0})||_2^2 \\ &\overset{\text{(i)}}{\leq} ||\mathbf{A}^{1/2}\mathbf{u}_{i_0}||_2 + ||\widehat{\mathbf{u}} - \mathbf{u}_{i_0}||_2 ||\mathbf{A}^{1/2}\widehat{\mathbf{u}}||_2 \\ &\overset{\text{(ii)}}{\leq} ||\mathbf{A}^{1/2}\mathbf{u}_{i_0}||_2 + \epsilon ||\mathbf{A}^{1/2}\widehat{\mathbf{u}}||_2, \end{aligned}$$

where inequality (i) is the triangle inequality on the ℓ_2-norm, and inequality (ii) follows from the definition of the ϵ-net. Thus, we get

$$\begin{aligned}\max_{\mathbf{u}\in\mathcal{S}^{n-1}} \mathbf{u}^\top \mathbf{A}\mathbf{u} &\leq \frac{1}{(1-\epsilon)^2}\mathbf{u}_{i_0}^\top \mathbf{A}\mathbf{u}_{i_0} \\ &\leq \frac{1}{(1-\epsilon)^2}\left(||\boldsymbol{\lambda}||_1 + \sqrt{2||\boldsymbol{\lambda}||_2 \cdot t} + 2||\boldsymbol{\lambda}||_\infty \cdot t\right).\end{aligned}$$

Noting that $\mu_{\mathsf{max}}(\mathbf{E}) = \mu_{\mathsf{max}}(\mathbf{A}) - ||\boldsymbol{\lambda}||_1$ gives us

$$\mu_{\mathsf{max}}(\mathbf{E}) \leq f_1(\boldsymbol{\lambda}; \epsilon, \tau). \tag{73}$$

On the other side, for any $\mathbf{u} \in \mathcal{S}^{n-1}$, let i^* be the index of its nearest neighbor in $\mathcal{U}$. Then, we have

$$\begin{aligned}\mathbf{u}^\top \mathbf{A}\mathbf{u} &= \mathbf{u}_{i^*}^\top \mathbf{A}\mathbf{u}_{i^*} + 2\mathbf{u}_{i^*}^\top \mathbf{A}(\mathbf{u}-\mathbf{u}_{i^*}) + (\mathbf{u}-\mathbf{u}_{i^*})^\top \mathbf{A}(\mathbf{u}-\mathbf{u}_{i^*}) \\ &\overset{\text{(i)}}{\geq} \mathbf{u}_{i^*}^\top \mathbf{A}\mathbf{u}_{i^*} + 2\mathbf{u}_{i^*}^\top \mathbf{A}(\mathbf{u}-\mathbf{u}_{i^*}) \\ &\overset{\text{(ii)}}{\geq} \mathbf{u}_{i^*}^\top \mathbf{A}\mathbf{u}_{i^*} - 2||\mathbf{A}^{1/2}\mathbf{u}_{i^*}||_2 \cdot ||\mathbf{A}^{1/2}(\mathbf{u}-\mathbf{u}_{i^*})||_2 \\ &\overset{\text{(iii)}}{\geq} \mathbf{u}_{i^*}^\top \mathbf{A}\mathbf{u}_{i^*} - 2||\mathbf{u}-\mathbf{u}_{i^*}||_2 \cdot ||\mathbf{A}^{1/2}\mathbf{u}_{i^*}||_2 \cdot ||\mathbf{A}^{1/2}\widehat{\mathbf{u}}||_2 \\ &\overset{\text{(iv)}}{\geq} \mathbf{u}_{i^*}^\top \mathbf{A}\mathbf{u}_{i^*} - 2\epsilon \cdot ||\mathbf{A}^{1/2}\mathbf{u}_{i^*}||_2 \cdot ||\mathbf{A}^{1/2}\widehat{\mathbf{u}}||_2 \\ &\geq \mathbf{u}_{i^*}^\top \mathbf{A}\mathbf{u}_{i^*} - \left(\frac{2\epsilon}{1-\epsilon}\right) \cdot ||\mathbf{A}^{1/2}\mathbf{u}_{i_0}||_2 \cdot ||\mathbf{A}^{1/2}\mathbf{u}_{i^*}||_2 \\ &\geq ||\boldsymbol{\lambda}||_1 - \sqrt{2||\boldsymbol{\lambda}||_2 \cdot t} - \left(\frac{2\epsilon}{1-\epsilon}\right) \cdot \left(||\boldsymbol{\lambda}||_1 + \sqrt{2||\boldsymbol{\lambda}||_2 \cdot t} + 2||\boldsymbol{\lambda}||_\infty \cdot t\right)\end{aligned}$$

where inequalities (i) and (ii) follow from the positive semidefiniteness of $\mathbf{A}$ and the Cauchy-Schwarz inequality respectively, and inequalities (iii) and (iv) follow from the definition of the operator norm and the ϵ-net respectively. The last two inequalities follow since we recall that $||\mathbf{A}^{1/2}\widehat{\mathbf{u}}||_2 \leq \frac{1}{(1-\epsilon)}||\mathbf{A}^{1/2}\mathbf{u}_{i_0}||_2$. Then, we substitute Equation (72a) twice for indices i^* and i_0 respectively. Again, noting that $\mu_{\min}(\mathbf{E}) = \mu_{\min}(\mathbf{A}) - ||\boldsymbol{\lambda}||_1$, and substituting $t := \tau + n\ln(1+\frac{2}{\epsilon})$, gives us

$$\mu_{\min}(\mathbf{E}) \geq -f_2(\boldsymbol{\lambda}; \epsilon, \tau). \tag{74}$$

Finally, using Equations (73) and (74), we have

$$||\mathbf{E}||_{\mathsf{op}} = \max\{\mu_{\max}(\mathbf{E}), -\mu_{\min}(\mathbf{E})\} \leq \max\{f_1(\boldsymbol{\lambda}; \epsilon, \tau), f_2(\boldsymbol{\lambda}; \epsilon, \tau)\},$$

completing the proof. ■

F.2 Proof of Lemma 21

In this subsection, we prove Lemma 21, i.e. concentration on the quantity $\frac{1}{\mathbf{u}^\top \mathbf{A}^{-1}\mathbf{u}}$ for the inverse Wishart matrix $\mathbf{A}^{-1}$. Because $\mathbf{A}$ is a Wishart matrix, we can use rotational

invariance of the distribution of the random variable $\mathbf{u}^\top \mathbf{A}^{-1}\mathbf{u}$ for any $\mathbf{u} \in \mathcal{S}^{n-1}$. Thus, it suffices to prove the concentration bound for $\mathbf{u} := \mathbf{e}_n$, i.e. study the random variable $\mathbf{A}^{-1}_{n,n} = \mathbf{e}_n^\top \mathbf{A}^{-1}\mathbf{e}_n$.

From elementary properties of the inverse Wishart distribution, we know that the quantity $\frac{1}{\mathbf{A}^{-1}_{n,n}} \sim \chi^2(d-n+1)$. Recall that we denoted $d'(n) := (d-n+1)$ for shorthand. Therefore, substituting Lemma 37 (with $\boldsymbol{\lambda} := \mathbf{1}$), we get

$$\Pr\left[\frac{1}{\mathbf{A}^{-1}_{n,n}} > \sqrt{2d'(n)t} + 2t\right] \leq e^{-t}$$
$$\Pr\left[\frac{1}{\mathbf{A}^{-1}_{n,n}} < -\sqrt{2d'(n)t}\right] \leq e^{-t}.$$

Since $\mathbf{A}^{-1}_{n,n}$ is identically distributed to $\mathbf{u}^\top \mathbf{A}^{-1}\mathbf{u}$ for any $\mathbf{u} \in \mathcal{S}^{n-1}$, the above concentration inequalities hold for the random variable $\frac{1}{\mathbf{u}^\top \mathbf{A}^{-1}\mathbf{u}}$. This completes the proof. ■

F.3 Proof of Lemma 27

In this subsection, we prove Lemma 27, i.e. equivalent quadratic form expressions for the contamination factor when binary labels are interpolated. As argued in Appendix D.3.1, for any $j \in \{1, \ldots, d\}$, the coefficient $\widehat{\alpha}_j$ is given by

$$\widehat{\alpha}_j = \mathbf{e}_j^\top \boldsymbol{\Phi}_{\mathsf{train}} \mathbf{A}^{-1} \mathbf{Y}_{\mathsf{train}} = \sqrt{\lambda_j}\mathbf{z}_j^\top \mathbf{A}^{-1}\mathbf{y}_t.$$

From the Sherman-Morrison-Woodbury identity, we have

$$\mathbf{A}^{-1} = \mathbf{A}_{-t}^{-1} - \frac{\lambda_t \mathbf{A}_{-t}^{-1}\mathbf{z}_t\mathbf{z}_t^\top \mathbf{A}_{-t}^{-1}}{1+\lambda_t \mathbf{z}_t^\top \mathbf{A}_{-t}^{-1}\mathbf{z}_t}.$$

Using this, we can rewrite $\widehat{\alpha}_j$ as

$$\begin{aligned}
\widehat{\alpha}_j &= \sqrt{\lambda_j}\mathbf{z}_j^\top \left(\mathbf{A}_{-t}^{-1} - \frac{\lambda_t \mathbf{A}_{-t}^{-1}\mathbf{z}_t\mathbf{z}_t^\top \mathbf{A}_{-t}^{-1}}{1+\lambda_t \mathbf{z}_t^\top \mathbf{A}_{-t}^{-1}\mathbf{z}_t}\right)\mathbf{y}_t \\
&= \sqrt{\lambda_j} \cdot \left(1 - \frac{1}{1+\lambda_t \mathbf{z}_t^\top \mathbf{A}_{-t}^{-1}\mathbf{z}_t}\right) \cdot \mathbf{z}_j^\top \mathbf{A}_{-t}^{-1}\mathbf{y}_t \\
&= \sqrt{\lambda_j} \cdot \mathbf{z}_j^\top \mathbf{A}_{-t}^{-1}\left(\mathbf{y}_t - \mathsf{SU}_b(t)\mathbf{z}_t\right)
\end{aligned}$$

where the last equality follows from Equation (47).
Using the definition of contamination (Equation (16)) and the above expressions, we get

$$\begin{aligned}
\mathsf{CN}_b^2(t) = \sum_{j=1, j\neq t}^{d} \lambda_j \widehat{\alpha}_j^2 &= \sum_{j=1, j\neq t}^{d} \lambda_j^2 \mathbf{y}_t^\top \mathbf{A}^{-1}\mathbf{z}_j\mathbf{z}_j^\top \mathbf{A}^{-1}\mathbf{y}_t \\
&= \mathbf{y}_t^\top \mathbf{A}^{-1}\left(\sum_{j=1, j\neq t}^{d} \lambda_j^2 \mathbf{z}_j\mathbf{z}_j^\top\right)\mathbf{A}^{-1}\mathbf{y}_t \\
&= \mathbf{y}_t^\top \mathbf{C}\mathbf{y}_t.
\end{aligned}$$

Now, we denote $\widetilde{\mathbf{y}}_t := \mathbf{y}_t - \mathsf{SU}_b(t)\mathbf{z}_t$. To prove the second form of contamination, we use the following sequence of equalities:

$$\begin{aligned}
\mathsf{CN}_b^2(t) = \sum_{j=1, j\neq t}^{d} \lambda_j \widehat{\alpha}_j^2 &= \sum_{j=1, j\neq t}^{d} \lambda_j \left(\sqrt{\lambda_j}\mathbf{z}_j^\top \mathbf{A}_{-t}^{-1}\widetilde{\mathbf{y}}_t\right)^2 \\
&= \sum_{j=1, j\neq t}^{d} \lambda_j^2 \widetilde{\mathbf{y}}_t{}^\top \mathbf{A}_{-t}^{-1}\mathbf{z}_j\mathbf{z}_j^\top \mathbf{A}_{-t}^{-1}\widetilde{\mathbf{y}}_t \\
&= \widetilde{\mathbf{y}}_t{}^\top \mathbf{A}_{-t}^{-1}\left(\sum_{j=1, j\neq t}^{d} \lambda_j^2 \mathbf{z}_j\mathbf{z}_j^\top\right)\mathbf{A}_{-t}^{-1}\widetilde{\mathbf{y}}_t \\
&= \widetilde{\mathbf{y}}_t{}^\top \widetilde{\mathbf{C}}\widetilde{\mathbf{y}}_t.
\end{aligned}$$

This completes the proof of Lemma 27. ■

F.4 Proof of Lemma 28

In this subsection, we prove Lemma 28, i.e. a high-probability upper bound on the quadratic forms $\widetilde{\mathbf{y}}_t{}^\top \widetilde{\mathbf{C}}\widetilde{\mathbf{y}}_t$ and $\mathbf{z}_t^\top \widetilde{\mathbf{C}}\mathbf{z}_t$ over only the randomness in $\{\mathbf{z}_t, \mathbf{y}_t\}$. Recall that we defined the random variables $\{\mathbf{z}_t, \mathbf{y}_t\}$ in Appendix D. Note that $\widetilde{\mathbf{C}}$ is almost surely positive definite and $\{\mathbf{z}_t, \widetilde{\mathbf{y}}_t\}$ are both pairwise independent of $\widetilde{\mathbf{C}}$. Further, note that

$$\begin{aligned}
\widetilde{\mathbf{y}}_t{}^\top \widetilde{\mathbf{C}}\widetilde{\mathbf{y}}_t &= (\mathbf{y}_t - \mathsf{SU}_b(t)\mathbf{z}_t)^\top \widetilde{\mathbf{C}}(\mathbf{y}_t - \mathsf{SU}_b(t)\mathbf{z}_t) \\
&\leq (\mathbf{y}_t - \mathsf{SU}_b(t)\mathbf{z}_t)^\top \widetilde{\mathbf{C}}(\mathbf{y}_t - \mathsf{SU}_b(t)\mathbf{z}_t) + (\mathbf{y}_t + \mathsf{SU}_b(t)\mathbf{z}_t)^\top \widetilde{\mathbf{C}}(\mathbf{y}_t + \mathsf{SU}_b(t)\mathbf{z}_t) \\
&= 2\mathbf{y}_t^\top \widetilde{\mathbf{C}}\mathbf{y}_t + 2\mathsf{SU}_b(t)^2 \mathbf{z}_t^\top \widetilde{\mathbf{C}}\mathbf{z}_t.
\end{aligned}$$

From Equation (44), we have

$$\mathbf{z}_t^\top \widetilde{\mathbf{C}}\mathbf{z}_t \leq \mathsf{tr}(\widetilde{\mathbf{C}})\left(1 + \frac{1}{c}\right) \cdot (\ln n)$$

almost surely for every realization of the random matrix $\widetilde{\mathbf{C}}$, and with probability at least $\left(1 - \frac{1}{n}\right)$ over the randomness in $\mathbf{z}_t$. By an identical argument (noting that $y_{t,i}^2 = 1$ almost surely, and that $\mathbb{E}\left[y_{t,i}y_{t,j}\right] = 0$ for any $i \neq j$), we can show that

$$\mathbf{z}_t^\top \widetilde{\mathbf{C}}\mathbf{z}_t \leq \mathsf{tr}(\widetilde{\mathbf{C}})\left(1 + \frac{1}{c}\right) \cdot (\ln n) \tag{75}$$

Substituting these inequalities in the expression for $\widetilde{\mathbf{y}}_t{}^\top \widetilde{\mathbf{C}}\widetilde{\mathbf{y}}_t$ completes the proof. ■

F.5 Proof of Lemma 30

In this subsection, we prove Lemma 30, i.e. a high-probability lower bound on the minimum eigenvalue of the random (almost surely positive semidefinite) matrix $\mathbf{C}$. Recall that we

defined

$$\mathbf{C} := \mathbf{A}^{-1}\left(\sum_{j=1,j\neq t}^{d}\lambda_j^2\mathbf{z}_j\mathbf{z}_j^\top\right)\mathbf{A}^{-1},$$
$$= \mathbf{A}^{-1}\left(\sum_{j=1}^{d-1}\widetilde{\lambda}_j^2\mathbf{z}_j\mathbf{z}_j^\top\right)\mathbf{A}^{-1}.$$

Using the mathematical fact from Appendix G.2, we have

$$\mu_n(\mathbf{C}) \geq (\mu_n(\mathbf{A}^{-1}))^2\mu_n\left(\sum_{j=1}^{d-1}\widetilde{\lambda}_j^2\mathbf{z}_j\mathbf{z}_j^\top\right).$$

Now, Equations (41) and (42a) from Lemma 24 can be used to lower bound the terms $(\mu_n(\mathbf{A}^{-1}))^2$ and $\mu_n\left(\sum_{j=1}^{d-1}\widetilde{\lambda}_j^2\mathbf{z}_j\mathbf{z}_j^\top\right)$ respectively. Substituting these lower bounds into the above bound completes the proof. ■

F.6 Proof of Lemma 31

In this subsection, we prove Lemma 31, i.e. equivalent quadratic form expressions for the contamination factor when real output is interpolated. This proof closely mirrors the proof of Lemma 27.

Let $\widehat{\alpha}_j$ denote the j^{th} component of $\widehat{\boldsymbol{\alpha}}_{2,\mathsf{real}}$. As argued in Appendix D.4.4, for any $j \in \{1,\ldots,d\}$, the coefficient $\widehat{\alpha}_j$ is given by

$$\widehat{\alpha}_j = \mathbf{e}_j^\top\boldsymbol{\Phi}_{\mathsf{train}}\mathbf{A}^{-1}\mathbf{Z}_{\mathsf{train}} = \sqrt{\lambda_j}\mathbf{z}_j^\top\mathbf{A}^{-1}\mathbf{z}_t. \tag{76}$$

By the Sherman-Morrison-Woodbury Identity, we have

$$\mathbf{A}^{-1} = \mathbf{A}_{-t}^{-1} - \frac{\lambda_t\mathbf{A}_{-t}^{-1}\mathbf{z}_t\mathbf{z}_t^\top\mathbf{A}_{-t}^{-1}}{1+\lambda_t\mathbf{z}_t^\top\mathbf{A}_{-t}^{-1}\mathbf{z}_t}.$$

Using this, we can rewrite $\widehat{\alpha}_j$ as

$$\widehat{\alpha}_j = \sqrt{\lambda_j}\left(1 - \frac{\lambda_t\mathbf{z}_t^\top\mathbf{A}_{-t}^{-1}\mathbf{z}_t}{1+\lambda_t\mathbf{z}_t^\top\mathbf{A}_{-t}^{-1}\mathbf{z}_t}\right)\mathbf{z}_j^\top\mathbf{A}_{-t}^{-1}\mathbf{z}_t$$
$$= \sqrt{\lambda_j}(1-\mathsf{SU}_r(t))\mathbf{z}_j^\top\mathbf{A}_{-t}^{-1}\mathbf{z}_t, \tag{77}$$

where the last equality follows from Equation (52).

Finally, using the definition of contamination (Equation (16)) together with Equation (76) gives us

$$\mathsf{CN}_r^2(t) = \sum_{j=1,j\neq t}^{d}\lambda_j\widehat{\alpha}_j^2 = \sum_{j=1,j\neq t}^{d}\lambda_j^2\mathbf{z}_t^\top\mathbf{A}^{-1}\mathbf{z}_j\mathbf{z}_j^\top\mathbf{A}^{-1}\mathbf{z}_t$$
$$= \mathbf{z}_t^\top\mathbf{A}^{-1}\left(\sum_{j=1,j\neq t}^{d}\lambda_j^2\mathbf{z}_j\mathbf{z}_j^\top\right)\mathbf{A}^{-1}\mathbf{z}_t$$
$$= \mathbf{z}_t^\top\mathbf{C}\mathbf{z}_t.$$

Similarly, applying Equation (77) gives us

$$\begin{aligned}
\mathsf{CN}_r^2(t) = \sum_{j=1, j\neq t}^{d} \lambda_j \widehat{\alpha}_j^2 &= \sum_{j=1, j\neq t}^{d} \lambda_j \left(\sqrt{\lambda_j}(1-\mathsf{SU}_r(t))\mathbf{z}_j^\top \mathbf{A}_{-t}^{-1}\mathbf{z}_t\right)^2 \\
&= (1-\mathsf{SU}_r(t))^2 \sum_{j=1, j\neq t}^{d} \lambda_j^2 \mathbf{z}_t^\top \mathbf{A}_{-t}^{-1}\mathbf{z}_j\mathbf{z}_j^\top \mathbf{A}_{-t}^{-1}\mathbf{z}_t \\
&= (1-\mathsf{SU}_r(t))^2 \mathbf{z}_t^\top \mathbf{A}_{-t}^{-1}\left(\sum_{j=1, j\neq t}^{d} \lambda_j^2 \mathbf{z}_j\mathbf{z}_j^\top\right)\mathbf{A}_{-t}^{-1}\mathbf{z}_t \\
&= (1-\mathsf{SU}_r(t))^2 \mathbf{z}_t^\top \widetilde{\mathbf{C}}\mathbf{z}_t.
\end{aligned}$$

This completes the proof. ■

Appendix G. Mathematical Facts

G.1 Upper bound on maximum eigenvalue of product of positive definite matrices

Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$ be symmetric positive definite matrices and let $\mathbf{C} = \mathbf{AB}$. It is a well known fact that for positive definite matrix $\mathbf{M}$, $\mu_1(\mathbf{M}) = \|\mathbf{M}\|_2$, i.e the largest eigenvalue is the operator norm. Using this,

$$\mu_1(\mathbf{C}) = \|\mathbf{C}\|_2 = \|\mathbf{AB}\|_2 \leq \|\mathbf{A}\|_2 \|\mathbf{B}\|_2 = \mu_1(\mathbf{A})\mu_1(\mathbf{B}),$$

where the inequality follows from the sub-multiplicativity of operator norm.

G.2 Lower bound on minimum eigenvalue of product of positive definite matrices

Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$ be symmetric positive definite matrices and let $\mathbf{C} = \mathbf{AB}$. Note that since inverses exist for positive definite matrices we can write,

$$\mu_n(\mathbf{C}) = \frac{1}{\mu_1(\mathbf{C}^{-1})} \geq \frac{1}{\mu_1(\mathbf{A}^{-1})\mu_1(\mathbf{B}^{-1})} = \mu_n(\mathbf{A})\mu_n(\mathbf{B}),$$

where the inequality follows by applying the upper bound for eigenvalue of product of two positive definite matrices from Appendix G.1.

Appendix H. Normalized margin calculations

In this section, we verify that the statement of Equation (19) exactly matches with the statement in Bartlett and Mendelson (2002, Theorem 21), using the notation from that paper. Observe that the first and third terms in the generalization bound exactly match. We only need to verify the second term. Note that the linear kernel is precisely

$$k(X, X') := \boldsymbol{\phi}(X)^\top \boldsymbol{\phi}(X').$$

Therefore, we get

$$\sqrt{\sum_{i=1}^{n} k(X_i, X_i)} = \sqrt{\sum_{i=1}^{n} ||\phi(X_i)||_2^2} \\ = ||\mathbf{\Phi}_{\mathsf{train}}||_{\mathsf{F}}.$$

Similarly, using the kernel trick (see the discussion just below Theorem 21 in the paper), we can verify that the term B is an upper bound on the quantity $||\widehat{\boldsymbol{\alpha}}||_2$. Substituting these equivalences into the original statement completes the verification.

References

Navid Azizan, Sahin Lale, and Babak Hassibi. A study of generalization of stochastic mirror descent algorithms on overparameterized nonlinear models. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3132–3136. IEEE, 2020.

Peter L Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE transactions on Information Theory*, 44(2):525–536, 1998.

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

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

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

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pages 6240–6249, 2017.

Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 2020. ISSN 0027-8424. doi: 10.1073/pnas.1907378117. URL `https://www.pnas.org/content/early/2020/04/22/1907378117`.

Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. *ICML*, 2018.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.

Shai Ben-David, David Loker, Nathan Srebro, and Karthik Sridharan. Minimizing the misclassification error rate using a surrogate convex loss. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 83–90, 2012.

Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, 1992.

Florentina Bunea. Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1 + \ell_2$ penalization. *Electronic Journal of Statistics*, 2:1153–1194, 2008.

Horen Chang. Presampling filtering, sampling and quantization effects on the digital matched filter performance. In *Proceedings of the International Telemetering Conference*, pages 889–915, 1982.

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

Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

Zeyu Deng, Abla Kammoun, and Christos Thrampoulidis. A model of double descent for high-dimensional binary linear classification. *arXiv preprint arXiv:1911.05822*, 2019.

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 1991.

Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.

Jianqing Fan and Jinchi Lv. Nonconcave penalized likelihood with np-dimensionality. *IEEE Transactions on Information Theory*, 57(8):5467–5484, 2011.

Jerome H Friedman. On bias, variance, 0/1—loss, and the curse-of-dimensionality. *Data mining and knowledge discovery*, 1(1):55–77, 1997.

Wei Gao and Zhi-Hua Zhou. On the doubt about margin explanation of boosting. *Artificial Intelligence*, 203:1–18, 2013.

Mario Geiger, Stefano Spigler, Stéphane d'Ascoli, Levent Sagun, Marco Baity-Jesi, Giulio Biroli, and Matthieu Wyart. Jamming transition as a paradigm to understand the loss landscape of deep neural networks. *Physical Review E*, 100(1):012115, 2019.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

Yehoram Gordon. Some inequalities for gaussian processes and applications. *Israel Journal of Mathematics*, 50(4):265–289, 1985.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pages 1832–1841, 2018.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics New York, 2 edition, 2009.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

Daniel Hsu, Vidya Muthukumar, and Ji Xu. On the proliferation of support vectors in high dimensions. In *International Conference on Artificial Intelligence and Statistics*, pages 91–99. PMLR, 2021.

Like Hui and Mikhail Belkin. Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks. In *9th International Conference on Learning Representations*, 2021.

Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Conference on Learning Theory*, pages 1772–1798, 2019.

Sham Kakade, Ohad Shamir, Karthik Sindharan, and Ambuj Tewari. Learning exponential families in high-dimensions: Strong convexity and sparsity. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 381–388, 2010.

Ganesh Ramachandra Kini and Christos Thrampoulidis. Analytic study of double descent in binary classification: The impact of loss. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2527–2532. IEEE, 2020.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

Yasaman Mahdaviyeh and Zacharie Naulet. Risk of the least squares minimum norm estimator under the spike covariance model. *arXiv*, pages arXiv–1912, 2019.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

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

Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.

Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 1(1):67–83, 2020.

Mor Shpigel Nacson, Jason Lee, Suriya Gunasekar, Pedro Henrique Pamplona Savarese, Nathan Srebro, and Daniel Soudry. Convergence of gradient descent on separable data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3420–3428, 2019.

Vaishnavh Nagarajan and J Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. In *Advances in Neural Information Processing Systems*, pages 11611–11622, 2019.

Preetum Nakkiran. More data can hurt for linear regression: Sample-wise double descent. *arXiv preprint arXiv:1912.07242*, December 2019.

Jeffrey Negrea, Gintare Karolina Dziugaite, and Daniel M Roy. In defense of uniform convergence: Generalization via derandomization with an application to interpolating predictors. *arXiv preprint arXiv:1912.04265*, 2019.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

Qichao Que and Mikhail Belkin. Back to the future: Radial basis function networks revisited. In Arthur Gretton and Christian C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 1375–1383, Cadiz, Spain, 2016. PMLR.

Ryan Michael Rifkin. *Everything old is new again: a fresh look at historical approaches in machine learning*. PhD thesis, Massachusetts Institute of Technology, 2002.

Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013.

Robert E Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. *The annals of statistics*, 26 (5):1651–1686, 1998.

Vatsal Shah, Anastasios Kyrillidis, and Sujay Sanghavi. Minimum norm solutions do not always generalize well for over-parameterized problems. *arxiv stat*, 1050:16, 2018.

C. Snyder and S. Vishwanath. Sample compression, support vectors, and generalization in deep learning. *IEEE Journal on Selected Areas in Information Theory*, 1(1):106–120, 2020.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

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

Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. Regularized linear regression: A precise analysis of the estimation error. In *Conference on Learning Theory*, pages 1683–1709, 2015.

George L. Turin. An introduction to digitial matched filters. *Proceedings of the IEEE*, 64 (7):1092–1112, 1976.

Sara A Van de Geer. High-dimensional generalized linear models and the LASSO. *The Annals of Statistics*, 36(2):614–645, 2008.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Weichen Wang and Jianqing Fan. Asymptotics of empirical eigenstructure for high dimensional spiked covariance. *Annals of statistics*, 45(3):1342, 2017.

Blake Woodworth, Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Kernel and deep regimes in overparametrized models. *arXiv preprint arXiv:1906.05827*, 2019.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, pages 56–85, 2004.