New Equivalences between Interpolation and SVMs: Kernels and Structured Features[*]

Chiraag Kaushik[†], Andrew D. McRae[‡], Mark Davenport[†], and Vidya Muthukumar[§]

Abstract. The support vector machine (SVM) is a supervised learning algorithm that finds a maximum-margin linear classifier, often after mapping the data to a high-dimensional feature space via the kernel trick. Recent work has demonstrated that in certain sufficiently overparameterized settings, the SVM decision function coincides exactly with the minimum-norm label interpolant. This phenomenon of support vector proliferation (SVP) is especially interesting because it allows us to understand SVM performance by leveraging recent analyses of harmless interpolation in linear and kernel models. However, previous work on SVP has made restrictive assumptions on the data/feature distribution and spectrum. In this paper, we present a new and flexible analysis framework for proving SVP in an arbitrary reproducing kernel Hilbert space with a flexible class of generative models for the labels. We present conditions for SVP for features in the families of general bounded orthonormal systems (e.g., Fourier features) and independent sub-Gaussian features. In both cases, we show that SVP occurs in many interesting settings not covered by prior work, and we leverage these results to prove novel generalization results for kernel SVM classification.

Key words. support vector machines, interpolation, overparameterization, kernel methods

MSC codes. 62H30, 62J07, 68Q32, 46E22

DOI. 10.1137/23M1568764

1. Introduction. Recent empirical and theoretical efforts in supervised machine learning have discovered a wide range of surprising phenomena that arise in the modern *overparameterized regime* (i.e., where the number of free parameters in the model is much larger than the number of training examples [13, 6]). For example, after it was observed that deep neural networks can perfectly fit noisy training data and still generalize well to new data (see, e.g., [35, 43]), several theoretical efforts demonstrated that this "harmless interpolation" phenomenon can in fact occur even in the simpler settings of linear and kernel regressions [8, 7, 4]. A separate, but equally surprising, observation in this overparameterized regime is that training

[†]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30308 USA (ckaushik7@gatech.edu, mdav@gatech.edu).
[‡]Institute of Mathematics, EPFL, Lausanne, Switzerland (andrew.mcrae@epfl.ch).
[§]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30308 USA, and School of Industrial & Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30308 USA (vmuthukumar8@gatech.edu).

procedures that optimize different loss functions can still yield similar test performance. For example, the empirical studies of [36, 22, 26, 16] demonstrate that kernel machines and deep neural networks trained using the squared loss, which is traditionally reserved for regression problems with continuous labels, can result in classification performance comparable to that of machines and networks trained with the more popular cross-entropy loss.

Motivated by these observations, recent work has sought to deepen theoretical understanding of the impact of the loss function in overparameterized classification tasks, starting with linear models. The hard-margin support vector machine (SVM) [12] is one popular classification algorithm which, given n linearly separable training samples $(x_i, y_i) \in \mathbf{R}^d \times \{-1, 1\}$, aims to find a decision function $\hat{f}(x) = \langle \hat{\theta}, x \rangle$ by solving the convex program

$$\hat{\theta} = \underset{\theta \in \mathbf{R}^d}{\arg\min} \|\theta\|_2 \ \text{ s.t. } \ y_i \theta^\top x_i \geq 1 \text{ for } i = 1, \ldots, n.$$

Equivalently, this solution maximizes the smallest distance from the hyperplane $\{x \colon \hat{f}(x) = 0\}$ to any training example. It was first shown in [33] that, in overparameterized settings, the SVM decision function can coincide exactly with the *minimum norm interpolant* of the training data (i.e., that the learned $\hat{\theta} = \arg\min_{\theta \in \mathbf{R}^d} \|\theta\|_2$ results in $\hat{f}(x_i) = y_i$ for all $i = 1, \ldots, n$). This phenomenon, termed *support vector proliferation (SVP)*, has several important implications. First, since it is known that these two solutions can be obtained via gradient descent on the squared and logistic losses, respectively, this gives theoretical support to the empirical observation that different loss functions can lead to similar behavior in overparameterized settings [36, 22]. This equivalence has also been exploited in several works (e.g., [33, 42, 41]) to study the *generalization* performance of the SVM by understanding the minimum-norm interpolant of the discrete labels, which can be more easily analyzed via a convenient closed-form solution. This approach can in fact be leveraged to prove classification consistency results for the SVM in settings which classical generalization bounds do not cover, as discussed in [33, 21].

Despite considerable subsequent progress in our understanding of SVP [21, 1, 42, 41, 10], current characterizations crucially depend on strong distributional assumptions on the data or features. Specifically, conditions under which SVP occurs have only been given in settings where the features have Gaussian or independent sub-Gaussian entries. Thus, these conditions are essentially applicable only when a linear model is learned directly on input data which is itself high-dimensional. By contrast, in the more general setting of *kernel SVM*, the training data $x \in \mathbf{R}^d$ is often first transformed (either explicitly or implicitly, using the kernel trick) via some feature map $\phi \colon \mathbf{R}^d \to \mathbf{R}^D$, and the SVM solution described above learns a function of the form $\hat{f}(x) = \langle \hat{\theta}, \phi(x) \rangle$. For essentially any choice of kernel or nonlinear feature map, the entries of $\phi(x_i)$ are *not* independent. Indeed, in many common learning settings, $D \gg d$; i.e., training data of *low or fixed* dimension (or effectively low, e.g., from a low-dimensional manifold) are mapped to some higher-dimensional or even infinite-dimensional features which are intricately dependent on one another. As a case study, SVP was first empirically observed in the experiments of [33], which used 1-dimensional data and a Fourier feature map whose entries have considerable structure. Existing theoretical results do not apply in such settings.

1.1. Contributions. In this paper, we study SVP for kernel methods (i.e., for general classes of feature maps) and provide precise nonasymptotic conditions under which it occurs.

We consider a generative label model for classification of the form $\eta^*(x) = 2\mathbf{P}(y = 1 \mid x) - 1$, where η^* is a function in an arbitrary RKHS (reproducing kernel Hilbert space) and represents a natural affine transformation of the classification conditional probability. Our results are briefly summarized below.

SVP for general bounded features. Under the label model described above, we provide sufficient conditions on the spectrum of the kernel integral operator and training examples under which the minimum RKHS-norm interpolant of the training data is identical to the decision function learned by the kernel SVM. In section 3.1 we formally provide the main assumptions under which our result holds. At a high level, we only require that the leading kernel eigenfunctions comprise a *bounded orthonormal system (BOS),* and we leverage recent conditions from the literature for harmless interpolation of noise to ensure that the minimum-norm interpolant is close to an *attenuated version*[1] of η^*. Then, in section 3.2 we characterize the settings in which SVP occurs under these assumptions. Compared to previous work, our requirements on the feature map and kernel spectrum are considerably weaker and allow us to show that SVP can occur in general RKHS settings without independence or sub-Gaussianity of features. Our proof techniques for these results depart significantly from prior literature that assumes worst-case labels [33, 21, 42, 41, 10, 1] and make novel conceptual connections between SVP and tools for sharp error analysis of interpolating methods. This allows us to leverage a few key insights from the harmless interpolation literature [31] to obtain refined bounds. Nevertheless, our proofs depart from these results in a few key respects. Unlike the L_2-norm bounds developed for generalization analysis, our approach relies on a combination of a "leave-one-out" technique (originating in [21]) and the derivation of fine-grained bounds on the *function evaluation* of an estimator at a specific point. Especially when features are not sub-Gaussian, this often requires careful handling and results in some unusual decompositions (e.g., Lemma 4.4) that do not follow from existing bounds in the kernel regression literature.

Refined bounds for independent, sub-Gaussian features. In section 3.3, we specialize and sharpen our results to the case where the features are independent and sub-Gaussian. Notably, unlike in prior works, our use of a probabilistic generative label model and weaker spectral assumptions allows us to show that SVP can occur even when the minimum-norm interpolant generalizes well for the regression task. To the best of our knowledge, this is the first instance in which SVP has been established for such cases.

Generalization of the kernel SVM in new regimes. As in previous work on SVP, our results can be combined with analysis of the minimum-norm interpolant to analyze the generalization performance of the SVM. In section 3.4, we show that our analysis can be used to prove asymptotic classification consistency of the kernel SVM in new settings which are too (effectively) high-dimensional for classical generalization bounds to be informative. Compared to prior works on SVP which consider linear models on the input data, our results demonstrate that the kernel SVM can generalize well for a range of interesting kernels that correspond to more structured feature mappings.

1.2. Related work. We organize our discussion of related work into two categories.

Support vector proliferation. The empirical observation that the squared and cross-entropy losses result in comparable classification performance in modern deep learning and kernel

---

[1]Specifically, we define an operator $\overline{\mathcal{S}}$ and require that $\hat{\eta} \approx \overline{\mathcal{S}}\eta^*$ and that $\overline{\mathcal{S}}$ be an L_∞ contraction operator.

models [36, 22, 26, 16] has led to increased theoretical interest in the phenomenon of support vector proliferation [33, 21, 1, 10, 42, 41] and, more generally, equivalences between the solutions obtained by various loss functions [27]. In particular, since the SVM and minimum-norm interpolant can be obtained by gradient descent on the cross-entropy (or similar exponential-tailed) [37, 24, 23, 18, 25] and squared losses, SVP provides an avenue for understanding the types of overparameterized settings in which this loss function equivalence can occur.

Prior works provide conditions (typically in terms of the data covariance and/or the signal-to-noise ratio) under which the phenomenon of SVP occurs with high probability, for an arbitrary (i.e., worst-case) choice of labels. However, these results all require strong assumptions on the data distribution and dimensionality. SVP was first shown under Gaussian design in [33], and sufficient conditions were given in [21] for independent sub-Gaussian and Haar design, along with a converse result for the Gaussian case. Furthermore, the authors of [1] provided a converse result for SVP for anisotropic sub-Gaussian data and proved that a data dimension $d = \Theta(n \log n)$ is needed for SVP in the isotropic case. We provide a more detailed contextualization between our upper bounds and the converse results of [1] in section 3.2.1. The work of [42, 10, 41] develops similar results under Gaussian and sub-Gaussian mixture models for the binary, multiclass, and one-versus-all SVM settings. In our work, we focus on a particular generative model for binary labels and do not explicitly make sub-Gaussian tail assumptions on the covariate distribution in our main result. A second key difference between our results and existing literature is that, because of our probabilistic generative model on the labels, we obtain results under weaker random matrix conditions (and hence conditions on the spectrum of the kernel integral operator). These spectral conditions are similar to those used in analyses of minimum-norm interpolants (e.g., [4, 34, 39, 32, 31]) and highlight a novel connection between SVP and harmless interpolation of noise.

Generalization analysis of classifiers in overparameterized settings. Classical bounds on the generalization error of the SVM rely on sample compression (e.g., [17, 15]) or Rademacher complexity (e.g., [5, 30]), which typically yields upper bounds on the test error in terms of the fraction of training samples which are support vectors or the margin achieved by the learned hyperplane on the training set, respectively. Moreover, the fraction of support vectors in the SVM was shown to be $o(1)$ in a range of settings where the number of training points is proportional to or much greater than the data dimension [14, 9, 29], implying good generalization. When the data or feature dimension is much larger than the number of training points, however, these results no longer apply. In fact, it has recently been shown that the SVM decision function can interpolate the training labels in highly overparameterized settings (i.e., that every point becomes a support vector) and that, as a result, traditional bounds become vacuous, in the sense that they are no better than a constant with high probability [33, 21].

Recent analyses of benign overfitting in linear and kernel regression models have provided an alternative path to studying the behavior and generalization of overparameterized classifiers. These results have been exploited[2] to provide error bounds in a signed Gaussian model [33], high-dimensional linear discriminant analysis [11], and Gaussian and sub-Gaussian

---

[2]We note that there is a wide body of work on asymptotic error analysis and general ℓ_p-regularizers and accordingly restrict our discussion to the most relevant setting, i.e., nonasymptotic error analyses of the SVM with ℓ_2-regularization.

mixture models [42, 41, 10]. Most of these works (with the exception of [11], which also requires very high-dimensional data) explicitly use SVP to link the performance of a classifier to that of an interpolating regression task, which is easier to analyze. Hence, these results only apply in the restrictive settings in which SVP has been shown to hold.

By contrast, our approach most closely resembles that of [31], which gives a bound for the classification loss of the label interpolant obtained by minimizing the squared loss in a general RKHS setting. The analysis tools provided in [31] in turn have some precedent in the approach of [20, 44], which considered explicitly regularized kernel regression under random design and minimal assumptions on the features. Notably, our results for bounded orthonormal systems are also stated in terms of spectral conditions on the kernel integral operator rather than particular properties of the data distribution. The resultant error expressions are not as sharp as those obtained via Gaussian or sub-Gaussian assumptions, but they apply for more general choices of kernel and feature map.

2. Background and setup. Our main results build on the analytic framework for kernel regression and classification in [31]. We summarize the main components of this framework here. Specifically, we introduce important concepts related to the spectral properties of reproducing kernel Hilbert spaces and set up the supervised learning framework under which we will study the equivalence of the support vector machine (SVM) and minimum-norm interpolation (MNI) classification problems.

2.1. Notation. For a probability space (X, μ) and $f \colon X \to \mathbf{R}$, we define the norms $\|f\|_{L_p} := (\mathbf{E}_{x \sim \mu} |f(x)|^p)^{1/p}$ and $\|f\|_{L_\infty} := \operatorname{ess\,sup}_x |f(x)|$. For $u \in \mathbf{R}^n$, $\|u\|_{\ell_2}$ or $\|u\|_2$ is the Euclidean norm. More generally, for a function f in a Hilbert space $\mathcal{H}$, $\|f\|_{\mathcal{H}}$ is the Hilbert norm. We define $L_2(X, \mu)$ to be the Hilbert space of square-integrable functions $f \colon X \to \mathbf{R}$ with inner product $\langle f, g \rangle_{L_2} = \mathbf{E}_{x \sim \mu} f(x) g(x)$. We denote the ℓ_2 and $\mathcal{H}$ inner products by $\langle \cdot, \cdot \rangle_{\ell_2}$ and $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, respectively. We will write the operator norm of an operator $T \colon \mathcal{H}_1 \to \mathcal{H}_2$ (for distinct Hilbert spaces $\mathcal{H}_1$, $\mathcal{H}_2$) with respect to the $\mathcal{H}_1$ and $\mathcal{H}_2$ norms as $\|T\|_{\mathcal{H}_1 \to \mathcal{H}_2} := \sup_{\|f\|_{\mathcal{H}_1} \leq 1} \|Tf\|_{\mathcal{H}_2}$ or, when $T \colon \mathcal{H}_1 \to \mathcal{H}_1$, simply as $\|T\|_{\mathcal{H}_1}$. For a subspace A of a Hilbert space $\mathcal{H}$, we denote the orthogonal projection operator onto A as $\mathcal{P}_A$, and for an operator T on $\mathcal{H}$, we denote $T_A := T\mathcal{P}_A$. Let $\operatorname{diag}^\perp(Z)$ denote the projection of a matrix Z onto the set of matrices with zero diagonal. The identity operator is denoted by $\mathcal{I}$ (or I when it is an ordinary matrix). The notation e_i denotes the ith canonical basis vector in $\mathbf{R}^n$. Finally, the notation $a \lesssim b$ means $a \leq Cb$ for some universal constant $C > 0$, $a \asymp b$ means $a \lesssim b$ and $b \lesssim a$, and $a \ll b$ means $a < cb$ for some (sufficiently small) constant $c > 0$.

2.2. Reproducing kernel Hilbert spaces. Let $\mathcal{H}$ be a reproducing kernel Hilbert space (RKHS) of real-valued functions on a probability space (X, μ). We denote the reproducing kernel of $\mathcal{H}$ as $k \colon X \times X \to \mathbf{R}$. By definition, k satisfies the property that, for any $x \in X$ and $f \in \mathcal{H}$, $f(x) = \langle k_x, f \rangle_{\mathcal{H}}$, and $k_x := k(x, \cdot)$ is the unique element in $\mathcal{H}$ with this property.

Our results rely crucially on the *spectral* properties of the kernel used during the learning process. To this end, we first assume that samples of X are drawn according to the probability measure μ. Then, the kernel integral operator $\mathcal{T} \colon L_2(X, \mu) \to L_2(X, \mu)$ is given by

$$\mathcal{T}f(\cdot) = \int_X k(\cdot, y) f(y) \, d\mu(y).$$

Assuming k is continuous, Mercer's theorem yields

$$k(x,y) = \sum_{\ell=1}^{\infty} \lambda_\ell v_\ell(x) v_\ell(y),$$

where $\{\lambda_\ell\}_{\ell=1}^{\infty}$ and $\{v_\ell\}_{\ell=1}^{\infty}$ are the eigenvalues and eigenfunctions of $\mathcal{T}$, respectively (a discussion of the relatively weak conditions under which such a decomposition holds can be found in [38]). Furthermore, the eigenfunctions $\{v_\ell\}_{\ell=1}^{\infty}$ form an orthonormal basis for $L_2(X, \mu)$. Without loss of generality, we will assume the eigenvalues are sorted so that $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$. For some index p, which we can select in our analysis, let $G := \mathrm{span}(v_1, \ldots, v_p)$ be the subspace of $L_2(X, \mu)$ corresponding to the largest p eigenvalues, and let $\mathcal{T}_G$ be the restriction of the kernel integral operator to G.

2.3. Classification setting. Assume we observe a training set $\{x_i, y_i\}_{i=1}^{n}$, consisting of i.i.d. (independent and identically distributed) copies of (x, y), where $x \sim \mu$ and $y \in \{-1, 1\}$. We assume the relationship between x and y is governed by the regression function

$$(2.1) \qquad \eta^*(x) := \mathbf{E}[y \mid x] = 2\mathbf{P}(y = 1 \mid x) - 1$$

for some $\eta^* \in G$.[3] In other words, given x, y equals 1 with probability $\frac{\eta^*(x)+1}{2}$ and equals -1 otherwise. Equation (2.1) constitutes a natural generative model for classification through performing regression on (an affine transformation of) the classification conditional probability; similar models appear in the literature on plug-in classifiers (e.g., [2]), albeit with slightly different regularity assumptions on η^*. Accordingly, we consider classification procedures which yield an estimator $\hat{\eta}$ of η^* and predict the label of a new point $x \in X$ via $\hat{y} = \mathrm{sign}(\hat{\eta}(x))$. It is often convenient to consider our binary observations as being of the form $y_i = \eta^*(x_i) + \xi_i$, where $\xi_i := y_i - \eta^*(x_i)$ is interpreted as observation noise and satisfies $\mathbf{E}[\xi \mid x] = 0$. We collect the observations y_i and the noise ξ_i into vectors y and ξ, respectively; it will be clear from the context whether these symbols refer to the scalar random variables or the vector of observed values.

Given a data set of the form described above, we let $\mathcal{A} : \mathcal{H} \to \mathbf{R}^n$ denote the sampling operator on $\mathcal{H}$, i.e., the operator such that $(\mathcal{A}f)_i = f(x_i)$. The adjoint $\mathcal{A}^* : \mathbf{R}^n \to \mathcal{H}$ of the sampling operator with respect to the $\mathcal{H}$ and ℓ_2 inner products is characterized by $\mathcal{A}^*\beta = \sum_{i=1}^{n} \beta_i k_{x_i}$, for any $\beta \in \mathbf{R}^n$. Finally, we denote the Gram matrix corresponding to the data set as $K := \mathcal{A}\mathcal{A}^* \in \mathbf{R}^{n \times n}$. The entries of this matrix are given by $K_{ij} = k(x_i, x_j)$.

For our analysis, we also split the sampling operator into its components which act on G and $G^\perp$, denoted $\mathcal{A}_G := \mathcal{C}$ and $\mathcal{R} := \mathcal{A}_{G^\perp}$, respectively. We let $\mathcal{C}^* = \mathcal{T}_G^{-1}\mathcal{A}_G^*$ be the adjoint operator of $\mathcal{A}_G$ corresponding to the inner product $\langle \cdot, \cdot \rangle_{L_2}$. Finally, the *excess classification risk* of an estimator $\hat{\eta}$ is defined as

$$\mathcal{E}(\hat{\eta}) := \mathbf{P}(\hat{y} \neq y) - \mathbf{P}(y \neq \mathrm{sign}(\eta^*)).$$

---

[3]The assumption that η^* lies in the subspace G is stated for convenience and to simplify the statement of our main assumption and theorems. It can be relaxed at the expense of an extra term in our error expressions. This is made explicit in section 5.

2.4. Ridgeless kernel regression. The ridgeless kernel regression, or **minimum-norm interpolation (MNI)**, procedure solves the problem

$$(2.2) \qquad \begin{aligned} \hat{\eta} &= \arg\min_{\eta \in \mathcal{H}} \|\eta\|_{\mathcal{H}} \\ &\text{s.t. } \eta(x_i) = y_i \text{ for all } i = 1, \ldots, n. \end{aligned}$$

This problem admits a closed-form solution $\hat{\eta}(x) = \sum_{i=1}^{n} \hat{\beta}_i k(x, x_i)$, where $\hat{\beta} = K^{-1} y$. The properties of such estimators have been studied extensively in several recent works (e.g., [28, 4, 19, 32, 31]). At a high level, most of these works decompose the learned estimate as

$$(2.3) \qquad \hat{\eta} = \mathcal{A}^*(\mathcal{A}\mathcal{A}^*)^{-1} y = \underbrace{\mathcal{A}^*(\mathcal{A}\mathcal{A}^*)^{-1} \mathcal{A}\eta^*}_{=:\hat{\eta}_0} + \underbrace{\mathcal{A}^*(\mathcal{A}\mathcal{A}^*)^{-1} \xi}_{=:\epsilon}$$

and then, under various assumptions, characterize when (i) the first term approximates η^* and (ii) the second term is negligible. Most central to our analysis, [31] shows that, under certain spectral conditions on the kernel, the estimate satisfies $\hat{\eta} \approx \overline{\mathcal{S}} \eta^*$, where $\overline{\mathcal{S}} = n\mathcal{T}_G \left(\overline{\alpha}\mathcal{I}_G + n\mathcal{T}_G\right)^{-1}$ is the *idealized survival operator*,[4] which represents how well the true function is recovered by MNI. We will see below that SVP can occur in a variety of settings where $\overline{\mathcal{S}}$ uniformly attenuates η^* in the L_∞-norm.

2.5. Kernel SVM. The **hard-margin kernel support vector machine (SVM)** procedure aims to find a linear classifier that maximizes the training data margin in the space $\mathcal{H}$. Particularly, the primal and dual forms of this problem can be written as

$$(2.4) \qquad \begin{aligned} \hat{\eta} &= \arg\min_{\eta \in \mathcal{H}} \|\eta\|_{\mathcal{H}} \\ &\text{s.t. } y_i \eta(x_i) \geq 1 \text{ for } i = 1, \ldots, n. \end{aligned} \qquad \Longleftrightarrow \qquad \begin{aligned} \hat{\beta} &= \arg\max_{\beta \in \mathbb{R}^n} y^\top \beta - \frac{1}{2} \beta^T K \beta \\ &\text{s.t. } y_i \beta_i \geq 0 \text{ for } i = 1, \ldots, n, \end{aligned}$$

with $\hat{\eta} = \sum_{i=1}^{n} \hat{\beta}_i k_{x_i}$. Observing the dual form, we note here that the estimators $\hat{\eta}$ obtained by the MNI and SVM learning procedures both have a representation as a linear combination (with coefficients given in the vector $\hat{\beta}$) of the input features k_{x_i}, for $i = 1, \ldots, n$. As was originally noted in the discussion of [33, 21], it is readily verified that the choice $\hat{\beta} = K^{-1} y$ (which characterizes the solution to (2.2)) also maximizes the dual objective function of (2.4); however, this choice may not satisfy all of the additional constraints of the SVM. Our main results in the following section aim to characterize when the estimators $\hat{\eta}$ obtained by solving these two problems coincide *exactly*. Since the *support vectors* are the training points which satisfy $\hat{\eta}(x_i) = \pm 1$ (and hence for which the primal constraints are active), this equivalence implies that every training point is a support vector in the learned SVM.

3. Main results. In this section, we provide our main results characterizing SVP in the RKHS setting described above. After describing our main assumptions and defining the key quantities which appear in our results, in section 3.2 we state a theorem for general bounded feature families and provide an example of a bilevel feature design which exhibits SVP. Then, in section 3.3 we use our analysis framework to strengthen our results in the case of independent,

---

[4]We will formally define $\overline{\alpha}$ in Assumption 3.1.

sub-Gaussian features, hence showing that SVP can occur in a more strictly expanded regime than that provided by the main results of [21] under our generative label model. Finally, we demonstrate that our result immediately implies good generalization (in the sense of statistical consistency) of the kernel SVM in new overparameterized settings.

3.1. Feature model and sampling conditions. We consider kernels for which the eigenfunctions of the kernel integral operator $v_\ell(x)$ satisfy one of the following two properties:

1. **Bounded orthonormal system on G (BOS):** The eigenfunctions of $\mathcal{T}_G$ satisfy $\sum_{\ell=1}^{p} v_\ell^2(x) \leq Cp$ with probability 1 over x for some constant $C \geq 1$.
2. **Independent sub-Gaussian (SG):** For $x \sim \mu$, the $\{v_\ell(x)\}_{\ell \geq 1}$ are independent, zero-mean, sub-Gaussian random variables.

The BOS condition on G is satisfied by several common feature families, such as Fourier features, certain orthonormal polynomial families, discrete orthonormal systems, etc. Notably, in many of these cases, the features are far from independent.

Our analysis considers the setting where the sampling locations in the training set are drawn independently from the distribution μ and satisfy certain properties with high probability (which can be separately verified via concentration bounds stated in Appendix A) The first two properties are equivalent to those in Theorem 1 of [31]. The last condition is a property of our analysis and is satisfied for a variety of choices of features and η^*. After listing our main assumption below, we provide a brief interpretation and discussion of when it is satisfied.

Assumption 3.1. For some $0 < \delta < 1/2$, the training samples $x_1, \ldots, x_n \overset{\text{i.i.d.}}{\sim} \mu$ satisfy all of the following with probability at least $1 - \delta$:

1. (Residual Gram matrix concentration.) There exist constants $\alpha_L \asymp \alpha_U > 0$ for which the residual Gram matrix satisfies

$$\alpha_L I_n \preccurlyeq \mathcal{R}\mathcal{R}^* \preccurlyeq \alpha_U I_n.$$

2. (Feature covariance concentration on G.) Let $\mathcal{C}_{\backslash i}$ be the sampling operator on G for the leave-one-out data set $\{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\}$ with L_2-adjoint $\mathcal{C}_{\backslash i}^*$. Then, for all $i \in \{1, \ldots, n\}$, the following holds for some $c \leq \frac{1}{2}$:

$$\frac{\alpha_U - \alpha_L}{\alpha_U + \alpha_L} + \frac{2}{n}\|\mathcal{C}_{\backslash i}^*\mathcal{C}_{\backslash i} - n\mathcal{I}_G\|_{L_2} \leq c.$$

3. (L_∞ control on expected recovered signal.) Let $\overline{\alpha}$ be the harmonic mean of α_L and α_U. Let $\overline{\mathcal{S}} := n\mathcal{T}_G(\overline{\alpha}\mathcal{I}_G + n\mathcal{T}_G)^{-1}$ be the idealized survival operator, and let $\overline{\mathcal{B}} := \mathcal{I}_G - \overline{\mathcal{S}}$ be the idealized bias operator. Then

$$b := \|\overline{\mathcal{B}}\eta^*\|_\infty \lesssim 1 - \|\overline{\mathcal{S}}\eta^*\|_\infty.$$

The first condition of residual Gram matrix concentration is needed to ensure harmless interpolation of noise in MNI and has been shown to hold with high probability in many situations, such as in the case of sub-Gaussian features [4, 39] and in settings where the leading eigenspace satisfies a hypercontractivity condition [32]. We provide relevant concentration bounds in Appendix A which can be used to verify this condition in both the BOS and SG

cases. Note that unlike previous results on SVP which are label-agnostic and depend crucially on concentration of the *entire* Gram matrix to a multiple of identity (including [33, 21]), we only require this concentration property for the residual component.

The second condition requires covariance concentration on G. The full-sample version of this assumption is standard in regression. Our leave-one-out version holds with high probability for BOS features when $n \gtrsim p \log n$ (see, e.g., Lemma A.3) and for SG features when $n \gtrsim p$ (Lemma A.4).

The third condition requires that the idealized survival operator $\overline{\mathcal{S}}$ shrinks η^* (approximately) *uniformly*. Although this condition seems rather unusual at first glance, it turns out to provide the flexibility necessary to cover previously studied regimes while also allowing us to show that SVP occurs in new settings (by leveraging the dependence on η^*). For instance, it is satisfied in any of the following cases:

- **Heavier-tailed eigenvalues:** In general, if $\overline{\alpha} \gg n\lambda_1$, then $\overline{\mathcal{S}}\eta^*$ will shrink towards 0 while $\overline{\mathcal{B}}\eta^*$ approaches η^*, and sufficient shrinkage coupled with bounded v_ℓ yields the desired L_∞ control. This condition is quite generic and covers most settings in which SVP has previously been studied—note the similarity to the condition $\sum_{i=1}^d \lambda_i \gg \lambda_1 n \log n$ from [21]. Special cases include slow polynomial decay (studied in the supplementary material (SVM_SIMODS_supplement.pdf [local/web 255KB])) and the isotropic finite-dimensional case (where $d \gg n$, $\lambda_\ell = 1$ for $\ell \leq d$, and 0 otherwise).
- **Identical leading eigenvalues:** The leading p eigenvalues of $\mathcal{T}$ are identically equal to some $\lambda > 0$. Then, $\overline{\mathcal{S}} = \frac{n\lambda}{\overline{\alpha}+n\lambda}\mathcal{I}_G$ and $\overline{\mathcal{B}} = \frac{\overline{\alpha}}{\overline{\alpha}+n\lambda}\mathcal{I}_G$, and the fact that $\|\eta^*\|_\infty \leq 1$ implies that the assumption holds (with the constant in the $\lesssim$ being exactly 1). We explore this scenario in section 3.2.1.
- **1-sparse target function:** We obtain the same result without identical leading eigenvalues if η^* is in the span of a single leading eigenvector, i.e., $\eta^* = \gamma v_\ell$ for some $1 \leq \ell \leq p$ and constant γ.

3.2. General conditions for SVM equivalence. Our main result provides sufficient conditions under Assumption 3.1 for every training point to be a support vector and hence for MNI and SVM to yield identical solutions. The proof of this result is given in section 4.

Theorem 3.2 (SVP in bounded orthonormal systems). *Consider the classification setting described in section 2 with BOS features. Suppose Assumption 3.1 is satisfied for $\eta^* \in G$. Then, with probability at least $1 - \delta - \frac{2}{n}$, the solutions of the MNI procedure (2.2) and the SVM procedure (2.4) coincide if*

$$(3.1) \qquad \frac{1}{b}\sqrt{\frac{p \log n}{n}} + \frac{\alpha_U - \alpha_L}{\alpha_U + \alpha_L}\left(\frac{\sqrt{\log n}}{b} + \sqrt{n}\right) + c^2\sqrt{p} \ll 1.$$

Our result depends, in a nuanced manner, on the idealized bias term b, the number of samples n, the dimension p of G, the concentration of our residual Gram matrix $\mathcal{R}\mathcal{R}^*$, and (through the constant c) the sample covariance concentration on G. This result shows that SVP can occur more generally in kernel settings without sub-Gaussian or independent features (including the Fourier feature setting in [33], where this phenomenon was originally observed). Note that, unlike the results of [33, 21], our analysis assumes a particular model for the labels

and does *not* allow for adversarial label selection. However, our results apply under much weaker assumptions on the features, as we discuss in the next section.

3.2.1. Implications for the bilevel ensemble. As a concrete example, we now introduce the overparameterized *bilevel ensemble* as defined in [33, 31]. This example is useful for understanding the types of settings in which SVP can occur and is parameterized by a tuple (n, β, q, r), where $\beta > 1$, $0 < r < 1$, and $q \leq \beta - r$. Here, $d = n^\beta$ is the number of features used for training, the first $p = n^r$ features correspond to larger eigenvalues, and

$$(3.2) \qquad \lambda_\ell = \begin{cases} 1 & \text{for } 1 \leq \ell \leq p, \\ n^{-\beta+r+q} & \text{for } p < \ell \leq d. \end{cases}$$

Note that $q \leq \beta - r$ (with equality representing the isotropic case) is required for correct eigenvalue ordering, and $r < 1$ ensures $p < n$. Here, q can be interpreted as controlling the "effective overparameterization," i.e., the separation between the leading and residual eigenvalues. This type of featurization was inspired by spiked covariance models [33] and has the useful property that it permits harmless interpolation of noise in linear regression throughout its permissible range of parameters, making it a natural setting in which to explore the phenomenon of SVP. While we specialize our results to this ensemble for better comparison to prior works, it is important to note that our main theorem is stated in considerable generality and holds for any feature map with relatively heavy-tailed eigenvalues (see also the discussion after Assumption 3.1). In the supplementary material (SVM_SIMODS_supplement.pdf [local/web 255KB]), we provide analogous results for the case of polynomial decay of eigenvalues (which was considered for benign overfitting in [4]) . Below we specialize the result of Theorem 3.2 to the case of bounded features under the bilevel ensemble. The proof is provided in Appendix B.

Corollary 3.3 (SVP in BOS: Bilevel ensemble). *Consider the bilevel ensemble with parameters (n, β, q, r) and bounded features $\{v_\ell(x)\}_{\ell=1}^d$. Suppose $\eta^* \in G$ as in the classification setting of section 2 and $\sup_x |\sum_{\ell>p} (|v_\ell(x)|^2 - 1)| \lesssim n^{\frac{\beta}{2}+1}$. Then, as $n \to \infty$, SVP occurs with probability $\to 1$ if $\beta > 3$, $q > \frac{1-r}{2}$, and $r < \frac{2}{3}$.*

This corollary demonstrates that SVP occurs in this setting as long as the degree of effective overparameterization (controlled by β and q) is sufficiently high. Notably, this permits situations where the features are highly structured. In fact, we only assume the v_ℓ are bounded and satisfy the thin-shell type condition $\sup_x \left| \sum_{\ell>p} (|v_\ell(x)|^2 - 1) \right| \ll n^{\frac{\beta}{2}+1}$, which ensures that $k(x, x)$ does not vary too much and is necessary for concentration of the residual Gram matrix (similar requirements are given in related works which study this concentration phenomenon, such as [32, 3]) This can be easily verified to hold for many structured feature families, including Fourier features, for which $|v_\ell(x)|^2 = 1$ uniformly for all x. We further note that, although we state asymptotic results under the bilevel model for clarity of exposition, Theorem 3.2 (and the proof in Appendix B) allows us to obtain finite-sample guarantees that hold with high probability in a wide range of settings with nonuniform eigenvalues.

3.3. Improvement in the independent, sub-Gaussian case. We now consider the case in which the features $\{v_\ell(x)\}_\ell$ are independent and sub-Gaussian random variables. This is

the setting which was considered in earlier works on support vector proliferation [21, 33]. These works generally do not assume a particular label model and state conditions in terms of certain notions of "effective rank" of the entire feature covariance. In a comparison to this line of work, we also state results in the independent, sub-Gaussian case. Our results apply to a more specific label model, which implicitly imposes an additional distributional assumption compared to previous work, since $\eta^*(x)$ (which is a linear combination of $\{v_i(x)\}_{i=1}^p$) must be uniformly bounded in our model. However, our analysis framework allows us to recover many existing results while also showing that SVP can occur in a broader range of settings. The proof of this theorem is given in section 4.5.

Theorem 3.4 (SVP for independent, sub-Gaussian features). *Consider the classification setting described in section* 2. *Suppose that Assumption* 3.1 *is satisfied for $\eta^* \in G$ and that the features $\{v_\ell(x)\}_{\ell>1}$ are zero-mean, independent, and sub-Gaussian. Then, with probability at least $1 - \delta - ne^{-p} - \frac{3}{n}$, the solutions of the MNI procedure* (2.2) *and the SVM procedure* (2.4) *coincide if*

$$(3.3) \qquad \frac{1}{b}\sqrt{\frac{p\log n}{n}} + \frac{\alpha_U - \alpha_L}{\alpha_U + \alpha_L}\frac{\sqrt{\log n}}{b} + \sqrt{\frac{\lambda_{p+1}n\log n}{\alpha_L}} + c\sqrt{\log n} \ll 1.$$

We can see that the last two terms are much sharper than the corresponding terms in the BOS case (e.g., $c\sqrt{\log n}$ instead of $c^2\sqrt{p}$). This improved bound, coupled with the faster concentration rates of the operators $\mathcal{RR}^*$ and $\mathcal{C}^*\mathcal{C}$ for sub-Gaussian features, can be shown to weaken the conditions for SVP under the bilevel ensemble. We state this condition in the following corollary, proved in Appendix B.

Corollary 3.5 (SVP for independent, sub-Gaussian case: Bilevel ensemble). *Consider the bilevel ensemble with parameters (n, β, q, r) and independent, zero-mean, sub-Gaussian features $\{v_\ell(x)\}_{\ell=1}^d$. Suppose $\eta^* \in G$ in the classification setting of section* 2. *Then, as $n \to \infty$, SVP occurs with probability $\to 1$ if $\beta > \max(1, 3 - 2r - 2q)$ and $q > \frac{1-r}{2}$.*

This corollary is consistent with the results of [21], which show that SVP occurs in the bilevel model for $q > 1 - r$ and $\beta > 1$ for arbitrary labels y_i. We recover this result under our label model while also finding that SVP can occur in the range $\frac{1-r}{2} < q < 1 - r$ as long as β is sufficiently large (i.e., as long as there is sufficient overparameterization). Unlike previous settings in which SVP was shown to occur, this is a regime in which the MNI procedure turns out to be regression-consistent, that is, it can recover η^* exactly in the limit as $n \to \infty$ [33].[5]

3.4. Generalization of the kernel SVM in new settings. In settings where support vector proliferation occurs, the equivalence between the MNI and the kernel SVM provides an approach for characterizing the generalization error of the SVM classifier. Specifically, we can use the MNI, which has a closed-form solution and has been analyzed extensively in recent literature, to understand the generalization of the kernel SVM. Indeed, as noted in [21], traditional SVM generalization bounds are often based on the fraction of support vectors and hence

---

[5]We note here that [1] showed that SVP is *unlikely* to happen (on sub-Gaussian data) for the regime where $d_2 := \frac{(\sum_{\ell \geq 1}\lambda_j)^2}{\sum_{\ell \geq 1}\lambda_j^2} \lesssim n\log n$ and $d_\infty^2 := \frac{(\sum_{j \geq 1}\lambda_j)^2}{\lambda_1^2} \gtrsim d_2 n$. However, this regime is not captured by any parameterization of the bilevel ensemble, as we always have $d_\infty^2 \ll d_2 n$. Our results apply to a different regime in which we allow $d_\infty \leq n$ and still show that SVP can hold.

become uninformative in the types of overparameterized settings where SVP occurs (and this fraction approaches 1). Nevertheless, the equivalence to MNI can be used to demonstrate that the kernel SVM can indeed generalize well (in the sense of asymptotic consistency) even in highly overparameterized settings where *every* training point is a support vector. Previously, such results were shown for linear models on Gaussian and independent-sub-Gaussian data [33, 21, 42, 41, 10]. We now demonstrate that such results can also be shown for general kernel methods under the bilevel ensemble, addressing in part an open question raised in [33]. Recall that under this ensemble, classical margin bounds can be used to show that the kernel SVM is classification-consistent for $q < 1 - r$. By leveraging generalization results for the MNI from [31], we show below that SVP can be used to show that the kernel SVM is also consistent in some cases where $q > 1 - r$, i.e., where there is a high degree of overparameterization.

Corollary 3.6 (generalization of kernel SVM in the bilevel ensemble). *Consider the bilevel ensemble with parameters (n, β, q, r). Suppose $\eta^* \in G$ in the classification setting of section 2. Then, the following hold:*

1. ***BOS case:*** *Under the conditions of Corollary 3.3, the excess classification risk of the kernel SVM satisfies*

$$\mathcal{E} \overset{\log}{\lesssim} \left(\max(1, n^{r+q-1}) \left(n^{(r-1)/2} + n^{(1-\beta)/2} \right) + n^{r+q-\beta/2} + n^{q-\frac{3}{2}(1-r)} \right) \|\eta^*\|_{L_2},$$

   with probability $\to 1$ as $n \to \infty$. Hence, $\mathcal{E} \to 0$ if $\beta > 3, r < \frac{2}{3}$, and $0 \le q < \frac{3}{2}(1-r)$.

2. ***SG case:*** *Under the conditions of Corollary 3.5, the excess classification risk of the kernel SVM satisfies*

$$\mathcal{E} \overset{\log}{\lesssim} \left(\max(1, n^{r+q-1}) \left(n^{(r-1)/2} + n^{(1-\beta)/2} \right) + n^{r+q-\frac{\beta}{2}-\frac{1}{2}} + n^{q-\frac{3}{2}(1-r)} \right) \|\eta^*\|_{L_2},$$

   with probability $\to 1$ as $n \to \infty$. Hence, $\mathcal{E} \to 0$ if $0 \le q < \min\left(1 - r + \frac{\beta-1}{2}, \frac{3}{2}(1-r)\right)$.

We note that the part 2 of Corollary 3.6 is very similar to the condition implied by the works [33, 21] while requiring a slightly stronger condition on q. However, these prior results assume that the true function is 1-sparse and linear; by contrast, we consider a flexible classification setting where the target function η^* is assumed to belong to the subspace G of $\mathcal{H}$, and the y_i follow the label model described in section 2. Furthermore, the part 1 of Corollary 3.6 applies to a broad range of nonindependent feature families with possibly heavier-tailed distributions. In the supplementary material (SVM_SIMODS_supplement.pdf [local/web 255KB]), we provide an analogous result for kernels with slow polynomial eigenvalue decay.

4. Proofs of main theorems. In this section, we provide proofs of our main results. We begin by describing our overall proof technique before considering the main components of the analysis separately, deferring the proofs of useful technical lemmas to Appendix A.

4.1. Overview of analysis. Our results are built off of the deterministic conditions for support vector proliferation provided in [21]. For self-containment, we first restate this result for the kernel setting with the notation introduced in section 2.

Proposition 4.1 (Lemma 1 in [21]). *For a kernel $k(x, y)$ corresponding to RKHS $\mathcal{H}$, let the minimum-norm label interpolant $\hat{\eta}$ be given by the formula*

$$\hat{\eta}(x) = \sum_{i=1}^{n} z_i k(x, x_i),$$

where $z = K^{-1}y$, and $K = [k(x_i, x_j)]_{ij}$ is the kernel Gram matrix, assumed to be nonsingular. Furthermore, let $\hat{\eta}_{\backslash i}$ be the minimum-norm label interpolant constructed using the leave-one-out data set consisting of points $\{(x_j, y_j)\}_{j=1, j\neq i}^{n}$.

Then, the following conditions are equivalent:

1. *Every data point (x_i, y_i) is a support vector for the max-margin SVM, i.e., the SVM boundary and the binary label interpolant are identical.*
2. *For every $i \in \{1, \dots, n\}$, $z_i y_i > 0$, i.e., $z_i \neq 0$ and $\mathrm{sign}(z_i) = y_i$.*
3. *For every $i \in \{1, \dots, n\}$, $y_i \hat{\eta}_{\backslash i}(x_i) < 1$.*

Our analysis aims to give conditions under which $|\hat{\eta}_{\backslash i}(x_i)| < 1$ for all i. When this occurs, Proposition 4.1 immediately implies that all training points are support vectors. Recalling the decomposition (2.3), we can write

$$|\hat{\eta}_{\backslash i}(x_i)| \leq |\epsilon_{\backslash i}(x_i)| + |\mathcal{P}_{G^{\perp}}\hat{\eta}_{\backslash i,0}(x_i)| + |\mathcal{P}_G \hat{\eta}_{\backslash i,0}(x_i) - \overline{\mathcal{S}}\eta^*(x_i)| + |\overline{\mathcal{S}}\eta^*(x_i)|.$$

By Assumption 3.1, $1 - |\overline{\mathcal{S}}\eta^*(x_i)| \gtrsim b$, so the above bound will be strictly less than 1 when

$$(4.1) \qquad \underbrace{\frac{|\epsilon_{\backslash i}(x_i)|}{b}}_{A} + \underbrace{\frac{|\mathcal{P}_{G^{\perp}}\hat{\eta}_{\backslash i,0}(x_i)|}{b}}_{B} + \underbrace{\frac{|\mathcal{P}_G \hat{\eta}_{\backslash i,0}(x_i) - \overline{\mathcal{S}}\eta^*(x_i)|}{b}}_{C} \lesssim 1$$

for all i. Term A represents the interpolation of label noise. Terms B and C represent how well the noiseless estimate $\hat{\eta}_{\backslash i,0}$ is approximated at x_i by the idealized survival $\overline{\mathcal{S}}\eta^*$: B represents the component orthogonal to G, and C is the component in G. In the following three subsections, which focus on the BOS feature case, we bound these three terms separately, yielding general conditions under which the condition (4.1) is satisfied (Theorem 3.2). Then, in section 4.5 we consider the case of independent, sub-Gaussian features and obtain sharper bounds for terms B and C (yielding Theorem 3.4).

4.2. Noise error. To bound term A in (4.1), first observe that the noise variables ξ are bounded with $|\xi_i| \leq 2$ for all $i \leq n$. Hence, the vector ξ is sub-Gaussian in the following sense: For any fixed vector $z \in \mathbf{R}^n$, $\langle z, \xi \rangle$ is a sub-Gaussian random variable with sub-Gaussian norm of $\|z\|_2$ within a constant. This means that, for any fixed $z \in \mathbf{R}^n$, with probability at least $1 - \delta'$ with respect to ξ, $|\langle z, \xi \rangle| \lesssim \|z\|_2 \sqrt{\log \delta'^{-1}}$. So, since

$$|\epsilon(x)| = |\langle k_x, \mathcal{A}^*(\mathcal{A}\mathcal{A}^*)^{-1}\xi \rangle_{\mathcal{H}}| = |\langle (\mathcal{A}\mathcal{A}^*)^{-1}\mathcal{A}k_x, \xi \rangle|,$$

it suffices to bound $\|(\mathcal{A}\mathcal{A}^*)^{-1}\mathcal{A}k_x\|_2$. We leverage this observation in the following lemma.

Lemma 4.2 (upper bound on the noise term). *Let $x \in X$ be independent of the training data, and suppose Assumption 3.1 holds. Then, with probability at least $1 - \frac{1}{n^2}$,*

$$|\epsilon(x)| \lesssim \sqrt{\frac{p \log n}{n}} + \frac{1}{\tilde{\alpha}} \|\mathcal{R}k_x^R\|_2 \sqrt{\log n},$$

where $\tilde{\alpha} = \frac{\alpha_L + \alpha_U}{2}$ and $k_x^R = \mathcal{P}_{G^{\perp}} k_x$.

Proof of Lemma 4.2. Let $x \in X$ be independent of the data. We have

$$\epsilon(x) = \langle k_x, \mathcal{A}^*(\mathcal{A}\mathcal{A}^*)^{-1}\xi \rangle_{\mathcal{H}} = \langle (\mathcal{A}\mathcal{A}^*)^{-1}\mathcal{A}k_x, \xi \rangle.$$

With respect to the randomness in ξ, this is a sub-Gaussian random variable whose sub-Gaussian norm is within a constant of $\|(\mathcal{A}\mathcal{A}^*)^{-1}\mathcal{A}k_x\|_2$. By Lemma 8 of [31], we have

$$\|(\mathcal{A}\mathcal{A}^*)^{-1}\mathcal{A}k_x\|_2 = \|(\mathcal{A}\mathcal{A}^*)^{-1}(\tilde{\alpha}I_n + \mathcal{A}_G\mathcal{A}_G^*)(\tilde{\alpha}I_n + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\mathcal{A}k_x\|_2$$

$$\leq \frac{1}{2}\left(\frac{\alpha_U}{\alpha_L} + 1\right)\|(\tilde{\alpha}I_n + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\mathcal{A}k_x\|_2$$

$$\lesssim \|(\tilde{\alpha}I_n + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\mathcal{A}_G k_x^G\|_2 + \|(\tilde{\alpha}I_n + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\mathcal{R}k_x^R\|_2,$$

where $\tilde{\alpha} = \frac{\alpha_L + \alpha_U}{2}$, and k_x^G and k_x^R denote $\mathcal{P}_G k_x$ and $\mathcal{P}_{G^\perp} k_x$, respectively. For the first of these terms, we have

$$\|(\tilde{\alpha}I_n + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\mathcal{A}_G k_x^G\|_2 = \|\mathcal{C}(\tilde{\alpha}\mathcal{I}_G + \mathcal{T}_G\mathcal{C}^*\mathcal{C})^{-1}k_x^G\|_2$$

$$= \|\mathcal{C}(\tilde{\alpha}\mathcal{T}_G^{-1} + \mathcal{C}^*\mathcal{C})^{-1}\mathcal{T}_G^{-1}k_x^G\|_2$$

$$\leq \|\mathcal{C}\|_{L_2 \to \ell_2}\|(\tilde{\alpha}\mathcal{T}_G^{-1} + \mathcal{C}^*\mathcal{C})^{-1}\|_{L_2 \to L_2}\|\mathcal{T}_G^{-1}k_x^G\|_{L_2}$$

$$\lesssim \sqrt{n}\left(\frac{1}{n}\right)\sqrt{p} = \sqrt{\frac{p}{n}}.$$

Here, the last inequality holds from the assumption on the concentration of $\mathcal{C}^*\mathcal{C}$ and the fact that $\|\mathcal{T}_G^{-1}k_x^G\|_{L_2}^2 = \|\sum_{\ell=1}^p v_\ell(x)v_\ell\|_{L_2}^2 = \sum_{\ell=1}^p v_\ell^2(x) \lesssim p$ by Parseval's identity and the BOS assumption. For the second term, we have

$$\|(\tilde{\alpha}I_n + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\mathcal{R}k_x^R\|_2 \leq \|(\tilde{\alpha}I_n + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\|\|\mathcal{R}k_x^R\|_2 \leq \frac{1}{\tilde{\alpha}}\|\mathcal{R}k_x^R\|_2.$$

Applying a sub-Gaussian tail bound yields the result.

We also note here that for independent, sub-Gaussian features, the above argument also applies, except here we use the fact that

$$\|\mathcal{T}_G^{-1}k_x^G\|_{L_2} = \left(\sum_{\ell=1}^p v_\ell^2(x)\right)^{1/2} \lesssim \sqrt{p}$$

with probability at least $1 - e^{-p}$ (see, e.g., [40, Theorem 3.1.1]); note that this inequality holds deterministically in the BOS case. So, the final result holds with probability at least $1 - e^{-p} - \frac{1}{n^2}$ rather than $1 - \frac{1}{n^2}$. We will use this fact in section 4.5. ∎

Now, we apply this lemma to the leave-one-out training set corresponding to index i and let $x = x_i$. In this case, $\|\mathcal{R}_{\backslash i}k_{x_i}^R\|_2 = \|\text{diag}^\perp(\mathcal{R}\mathcal{R}^*)e_i\|_2 \leq \|(\mathcal{R}\mathcal{R}^* - \alpha_L I_n)e_i\|_2 \leq \alpha_U - \alpha_L$. Recalling the definition of $\tilde{\alpha}$, this yields

$$|\epsilon_{\backslash i}(x_i)| \lesssim \sqrt{\frac{p \log n}{n}} + \frac{\alpha_U - \alpha_L}{\alpha_U + \alpha_L}\sqrt{\log n}$$

with probability at least $1 - \frac{1}{n^2}$. We can then take a union bound over the n leave-one-out sets and divide by b to obtain the final bound on term A: with probability at least $1 - \frac{1}{n}$, for all $i = 1, \ldots, n$,

$$(4.2) \qquad A \lesssim \frac{1}{b} \cdot \sqrt{\frac{p \log n}{n}} + \frac{1}{b} \cdot \frac{\alpha_U - \alpha_L}{\alpha_U + \alpha_L}\sqrt{\log n}.$$

4.3. Bias error due to $G^\perp$. For term B in (4.1),

$$|\mathcal{P}_{G^\perp}\hat{\eta}_{\backslash i,0}(x_i)| = \left|\sum_{j=1,j\neq i}^{n}(K_{\backslash i}^{-1}\mathcal{A}_{\backslash i}\eta^*)_j k^R(x_j,x_i)\right| \leq \left\|K_{\backslash i}^{-1}\mathcal{A}_{\backslash i}\eta^*\right\|_2 \left\|\mathrm{diag}^\perp(\mathcal{R}\mathcal{R}^*)e_i\right\|_2$$

by the Cauchy–Schwarz inequality. The following lemma bounds the first of these terms.

Lemma 4.3 (useful bound for $G^\perp$ term). *Under Assumption 3.1 and the classification model described in section 2, $\|K^{-1}\mathcal{A}\eta^*\|_2 \lesssim \frac{b\sqrt{n}}{\alpha_L}$.*

Proof of Lemma 4.3. Recall the decomposition $K = \mathcal{A}_G\mathcal{A}_G^* + \mathcal{R}\mathcal{R}^*$. Then, by the push-through identity,

$$
\begin{aligned}
\|(\mathcal{R}\mathcal{R}^* + \mathcal{A}_G\mathcal{A}_G)^{-1}\mathcal{A}\eta^*\|_2 &= \|(\mathcal{R}\mathcal{R}^*)^{-1}\mathcal{A}_G(\mathcal{I}_G + \mathcal{A}_G^*(\mathcal{R}\mathcal{R}^*)^{-1}\mathcal{A}_G)^{-1}\eta^*\|_2 \\
&= \|(\mathcal{R}\mathcal{R}^*)^{-1}\mathcal{A}_G\mathcal{B}\eta^*\|_2 \\
&\leq \|(\mathcal{R}\mathcal{R}^*)^{-1}\|_{\ell_2\to\ell_2}\|\mathcal{A}_G\|_{L_2\to\ell_2}\|\mathcal{B}\eta^*\|_{L_2} \\
&\lesssim \frac{1}{\alpha_L}\sqrt{n}b\left(1 + \frac{c}{1-c}\right) \lesssim \frac{b\sqrt{n}}{\alpha_L}.
\end{aligned}
$$

In the last line, we use the fact that $\|\mathcal{A}_G\|_{L_2\to\ell_2} = \sqrt{\|\mathcal{C}^*\mathcal{C}\|_{L_2}} \asymp \sqrt{n}$ by part 2 of Assumption 3.1. Furthermore, we have denoted $\mathcal{B} := (\mathcal{I}_G + \mathcal{A}_G^*(\mathcal{R}\mathcal{R}^*)^{-1}\mathcal{A}_G)^{-1}$ and used Lemma A.5 to bound this "bias operator" by the "idealized bias operator" defined in Assumption 3.1: $\|\mathcal{B}\eta^*\|_{L_2} \leq (1 + \frac{c}{1-c})\|\overline{\mathcal{B}}\eta^*\|_{L_2} \leq (1 + \frac{c}{1-c})b$. ∎

Using this lemma, noting (as in the previous section) that $\left\|\mathrm{diag}^\perp(\mathcal{R}\mathcal{R}^*)e_i\right\|_2 \leq \alpha_U - \alpha_L$, and dividing by b, we conclude that term B is bounded (within a constant) by $(\frac{\alpha_U}{\alpha_L} - 1)\sqrt{n} \asymp \frac{\alpha_U - \alpha_L}{\alpha_U + \alpha_L}\sqrt{n}$.

4.4. Bias error in G. For term C, we first note that $|\mathcal{P}_G\hat{\eta}_{\backslash i,0}(x_i) - \overline{\mathcal{S}}\eta^*(x_i)|$ corresponds to the absolute error between the predictions of the G-projected estimator and its "idealized" version, given by $\overline{\mathcal{S}}\eta^*$. In Lemma 4.4, we prove a high probability bound for $|\mathcal{P}_G\hat{\eta}_0(x) - \overline{\mathcal{S}}\eta^*(x)|$, where x is independent of all training points.

Lemma 4.4 (error of the G-projected estimator). *Let $x \in X$ be independent of the training data, and suppose Assumption 3.1 holds. Then, with probability at least $1 - \frac{1}{n^2}$,*

$$\frac{1}{b}|\mathcal{P}_G\hat{\eta}_0(x) - \overline{\mathcal{S}}\eta^*(x)| \lesssim \sqrt{\frac{p\log n}{n}} + \sqrt{p}\frac{\alpha_U - \alpha_L}{\alpha_U + \alpha_L} + c^2\sqrt{p}.$$

Proof of Lemma 4.4. First, note that we can write $|\mathcal{P}_G\hat{\eta}_0(x) - \overline{\mathcal{S}}\eta^*(x)| = |((\mathcal{B} - \overline{\mathcal{B}})\eta^*)(x)|$, where $\mathcal{B} = (\mathcal{I}_G + \mathcal{A}_G^*(\mathcal{R}\mathcal{R}^*)^{-1}\mathcal{A}_G)^{-1} =: (\mathcal{I}_G + \mathcal{M})^{-1}$ and $\overline{\mathcal{B}} = (\mathcal{I}_G + \frac{n}{\alpha}\mathcal{T}_G)^{-1} =: (\mathcal{I}_G + \overline{\mathcal{M}})^{-1}$. Then note that

$$\mathcal{B} - \overline{\mathcal{B}} = \overline{\mathcal{B}}(\overline{\mathcal{M}} - \mathcal{M})\mathcal{B} = \overline{\mathcal{B}}(\overline{\mathcal{M}} - \mathcal{M})\overline{\mathcal{B}} + \overline{\mathcal{B}}(\overline{\mathcal{M}} - \mathcal{M})(\mathcal{B} - \overline{\mathcal{B}}).$$

So, we get

$$
\begin{aligned}
|\mathcal{P}_G\hat{\eta}_0(x) - \overline{\mathcal{S}}\eta^*(x)| &\leq |(\overline{\mathcal{B}}(\overline{\mathcal{M}} - \mathcal{M})\overline{\mathcal{B}}\eta^*)(x)| + |(\overline{\mathcal{B}}(\overline{\mathcal{M}} - \mathcal{M})(\mathcal{B} - \overline{\mathcal{B}})\eta^*)(x)| \\
&\leq |(\overline{\mathcal{B}}(\overline{\mathcal{M}} - \mathcal{M})\overline{\mathcal{B}}\eta^*)(x)| + \|\overline{\mathcal{B}}(\overline{\mathcal{M}} - \mathcal{M})(\mathcal{B} - \overline{\mathcal{B}})\eta^*\|_{L_\infty}.
\end{aligned}
$$

We then use the decomposition

$$\overline{\mathcal{M}} - \mathcal{M} = \frac{n}{\overline{\alpha}}\mathcal{T}_G - \mathcal{A}_G^*(\mathcal{R}\mathcal{R}^*)^{-1}\mathcal{A}_G = \frac{n}{\overline{\alpha}}\left(\mathcal{T}_G - \frac{1}{n}\mathcal{A}_G^*\mathcal{A}_G\right) + \mathcal{A}_G^*\left(\frac{1}{\overline{\alpha}}\mathcal{I}_n - (\mathcal{R}\mathcal{R}^*)^{-1}\right)\mathcal{A}_G,$$

which, applied to the first term, yields

(4.3)

$$
\begin{aligned}
|\mathcal{P}_G\hat{\eta}_0(x) - \overline{\mathcal{S}}\eta^*(x)| &\leq \left|\left(\overline{\mathcal{B}}\frac{n}{\overline{\alpha}}\left(\mathcal{T}_G - \frac{1}{n}\mathcal{A}_G^*\mathcal{A}_G\right)\overline{\mathcal{B}}\eta^*\right)(x)\right| \\
&\quad + \left|\left(\overline{\mathcal{B}}\mathcal{A}_G^*\left(\frac{1}{\overline{\alpha}}\mathcal{I}_n - (\mathcal{R}\mathcal{R}^*)^{-1}\right)\mathcal{A}_G\overline{\mathcal{B}}\eta^*\right)(x)\right| + \left\|\overline{\mathcal{B}}(\overline{\mathcal{M}} - \mathcal{M})(\mathcal{B} - \overline{\mathcal{B}})\eta^*\right\|_{L_\infty} \\
&\leq \left|\left(\overline{\mathcal{B}}\frac{n}{\overline{\alpha}}\left(\mathcal{T}_G - \frac{1}{n}\mathcal{A}_G^*\mathcal{A}_G\right)\overline{\mathcal{B}}\eta^*\right)(x)\right| \\
&\quad + \left\|\overline{\mathcal{B}}\mathcal{A}_G^*\left(\frac{1}{\overline{\alpha}}\mathcal{I}_n - (\mathcal{R}\mathcal{R}^*)^{-1}\right)\mathcal{A}_G\overline{\mathcal{B}}\eta^*\right\|_{L_\infty} + \left\|\overline{\mathcal{B}}(\overline{\mathcal{M}} - \mathcal{M})(\mathcal{B} - \overline{\mathcal{B}})\eta^*\right\|_{L_\infty}.
\end{aligned}
$$

We obtain the desired bound by separately bounding each of the three terms in (4.3). For the first term, note that

$$
\begin{aligned}
\overline{\mathcal{B}} \cdot \frac{n}{\overline{\alpha}}\left(\mathcal{T}_G - \frac{1}{n}\mathcal{A}_G^*\mathcal{A}_G\right)\overline{\mathcal{B}} &= \frac{n}{\overline{\alpha}}\left(\mathcal{I}_G + \frac{n}{\overline{\alpha}}\mathcal{T}_G\right)^{-1}\left(\mathcal{T}_G - \frac{1}{n}\mathcal{A}_G^*\mathcal{A}_G\right)\overline{\mathcal{B}} \\
&= \left(\frac{\overline{\alpha}}{n}\mathcal{T}_G^{-1} + \mathcal{I}_G\right)^{-1}\left(\mathcal{I}_G - \frac{1}{n}\mathcal{C}^*\mathcal{C}\right)\overline{\mathcal{B}}.
\end{aligned}
$$

Rather than bounding an L_∞ norm, we directly analyze the application of this operator to η^*. Let $\mathcal{D} := \left(\frac{\overline{\alpha}}{n}\mathcal{T}_G^{-1} + \mathcal{I}_G\right)^{-1}$; note that $\mathcal{D}$ is a contraction on G (i.e., $\|\mathcal{D}f\|_{L_2} \leq \|f\|_{L_2}$ for $f \in G$). Furthermore, let $\delta_{x_i}^G = \mathcal{T}_G^{-1}k_{x_i}^G$. Then, for fixed $f \in G$ and $x \in X$, we have

$$\left|\left(\mathcal{D}\left(\mathcal{I}_G - \frac{1}{n}\mathcal{C}^*\mathcal{C}\right)f\right)(x)\right| = \left|\frac{1}{n}\sum_{i=1}^n (f(x_i)\mathcal{D}\delta_{x_i}^G(x) - (\mathcal{D}f)(x))\right|,$$

which is a sum of independent zero-mean random variables. Specifically, each of the variables $f(x_i)(\mathcal{D}\delta_{x_i}^G)(x) - (\mathcal{D}f)(x)$ is zero-mean and has variance

$$\mathbf{E}_{x_i}|f(x_i)(\mathcal{D}\delta_{x_i}^G)(x) - (\mathcal{D}f)(x)|^2 \leq \mathbf{E}_{x_i}|f(x_i)(\mathcal{D}\delta_{x_i}^G)(x)|^2 \leq \|f\|_\infty^2\|\mathcal{D}\delta_x^G\|_{L_2}^2 \lesssim p\|f\|_\infty^2,$$

where we use the fact that D is a contraction and the BOS assumption in the last inequality. Moreover, each term is bounded as $|f(x_i)(\mathcal{D}\delta_{x_i}^G)(x) - (\mathcal{D}f)(x)| \lesssim \|f\|_\infty p$. So, Bernstein's inequality (e.g., [40, section 2.8]) gives, with probability at least $1 - \frac{1}{n^2}$,

$$\left|\left(\mathcal{D}\left(\mathcal{I}_G - \frac{1}{n}\mathcal{C}^*\mathcal{C}\right)f\right)(x)\right| \lesssim \sqrt{\frac{p\log n}{n}}\|f\|_\infty + \frac{p\log n}{n}\|f\|_\infty \lesssim \sqrt{\frac{p\log n}{n}}\|f\|_\infty.$$

We bound the first term by applying this with $f = \overline{\mathcal{B}}\eta^*$ and recalling that $\|\overline{\mathcal{B}}\eta^*\|_\infty = b$.

For the second term of (4.3), it is proved in Lemma 7 of [31] (which we state as Lemma A.5) that

$$\left\|\overline{\mathcal{B}}\mathcal{A}_G^*\left(\frac{1}{\overline{\alpha}}I_n - (\mathcal{R}\mathcal{R}^*)^{-1}\right)\mathcal{A}_G\right\|_{L_2} \leq \frac{\alpha_U - \alpha_L}{\alpha_U + \alpha_L}\left(1 + \frac{1}{n}\|\mathcal{C}^*\mathcal{C} - n\mathcal{I}_G\|_{L_2}\right) \asymp \frac{\alpha_U - \alpha_L}{\alpha_U + \alpha_L},$$

where the last expression follows by the part 2 of Assumption 3.1. Recall that, by the BOS assumption, for all $f \in G$, $\|f\|_{L_\infty} \lesssim \sqrt{p}\|f\|_{L_2}$. So, applying this to $f = \overline{\mathcal{B}}\eta^*$ and noting that $\|\overline{\mathcal{B}}\eta^*\|_{L_2} \leq b$, we get

$$\left\|\overline{\mathcal{B}}\mathcal{A}_G^*\left(\frac{1}{\overline{\alpha}}I_n - (\mathcal{R}\mathcal{R}^*)^{-1}\right)\mathcal{A}_G\overline{\mathcal{B}}\eta^*\right\|_{L_\infty} \lesssim \sqrt{p}\frac{\alpha_U - \alpha_L}{\alpha_U + \alpha_L}b.$$

To bound the third term, note that $\overline{\mathcal{B}}(\overline{\mathcal{M}} - \mathcal{M})$ is precisely the operator $\mathcal{Q}$ from Lemma A.5. Hence, $\|\overline{\mathcal{B}}(\overline{\mathcal{M}} - \mathcal{M})\|_{L_2} \leq c$. Furthermore, we have (again using the result of Lemma A.5)

$$(4.4) \qquad \|(\overline{\mathcal{B}} - \mathcal{B})\eta^*\|_{L_2} \leq \frac{c}{1-c}\|\overline{\mathcal{B}}\eta^*\|_{L_2} \leq \frac{c}{1-c}\|\overline{\mathcal{B}}\eta^*\|_{L_\infty} \leq \frac{bc}{1-c}.$$

Now, using the BOS assumption, the third term in (4.3) can be bounded as

$$\|(\overline{\mathcal{B}}(\overline{\mathcal{M}} - \mathcal{M})(\mathcal{B} - \overline{\mathcal{B}})\eta^*\|_{L_\infty} \leq \sqrt{p}\cdot\|\overline{\mathcal{B}}(\overline{\mathcal{M}} - \mathcal{M})\|_{L_2}\|(\overline{\mathcal{B}} - \mathcal{B})\eta^*\|_{L_2} \leq \frac{bc^2\sqrt{p}}{1-c} \lesssim bc^2\sqrt{p}.$$

Combining all of the above bounds, we finally arrive at

$$|\mathcal{P}_G\hat{\eta}_0(x) - \overline{\mathcal{S}}\eta^*(x)| \lesssim \sqrt{\frac{p\log n}{n}}\cdot b + \frac{\alpha_U - \alpha_L}{\alpha_U + \alpha_L}\cdot b\sqrt{p} + bc^2\sqrt{p}$$

with probability at least $1 - \frac{1}{n^2}$. ■

To bound term C in (4.1), we can apply this lemma to the n leave-one-out training sets and take a union bound to get the final bound, with probability at least $1 - \frac{1}{n}$. Combining the bounds on terms A, B, and C in the previous three subsections yields Theorem 3.2.

4.5. Proofs for sub-Gaussian features. Suppose that instead of satisfying the BOS condition, the features $v_\ell(x)$ are sub-Gaussian and independent. In this case, we first note that for all $f \in \mathcal{H}$, $f(x) = \langle k_x, f\rangle_{\mathcal{H}}$ is a sub-Gaussian random variable with norm at most $\|f\|_{L_2}$, up to constants. Using this observation, we can obtain sharp bounds for each term in (4.1).

For term A, we note that the statement of Lemma 4.2 also holds in the independent, sub-Gaussian case, with probability at least $1 - \frac{1}{n^2} - e^{-p}$ (this fact is verified in the proof of the lemma). So, after applying Lemma 4.2 to each of the n leave-one-out sets, we obtain the same upper bound for term A as in the BOS case, with probability $1 - ne^{-p} - \frac{1}{n}$.

For term B we note that the term $|\mathcal{P}_{G^\perp}\hat{\eta}_{\setminus i,0}(x)|$ is sub-Gaussian with norm at most $\|\mathcal{P}_{G^\perp}\mathcal{A}^*(\mathcal{A}\mathcal{A}^*)^{-1}\mathcal{A}\eta^*\|_{L_2}$. We bound this in the following lemma.

Lemma 4.5 ($G^\perp$ error term for sub-Gaussian features). *Suppose Assumption* 3.1 *holds for independent, sub-Gaussian features. Then,*

$$\left\|\mathcal{P}_{G^\perp}\mathcal{A}^*(\mathcal{A}\mathcal{A}^*)^{-1}\mathcal{A}\eta^*\right\|_{L_2} \lesssim \sqrt{\frac{n\lambda_{p+1}}{\alpha_L}}b.$$

Proof of Lemma 4.5. We proceed in a similarly to the proof of Lemma 4.3.

$$
\begin{aligned}
\left\|\mathcal{P}_{G^\perp}\mathcal{A}^*(\mathcal{A}\mathcal{A}^*)^{-1}\mathcal{A}\eta^*\right\|_{L_2} &\le \|\mathcal{I}_{G^\perp}\|_{\mathcal{H}\to L_2}\|\mathcal{R}^*(\mathcal{A}\mathcal{A}^*)^{-1}\mathcal{A}\eta^*\|_{\mathcal{H}} \\
&= \|\mathcal{I}_{G^\perp}\|_{\mathcal{H}\to L_2}\|\mathcal{R}^*(\mathcal{R}\mathcal{R}^* + \mathcal{A}_G\mathcal{A}_G^*)^{-1}\mathcal{A}\eta^*\|_{\mathcal{H}} \\
&= \|\mathcal{I}_{G^\perp}\|_{\mathcal{H}\to L_2}\|\mathcal{R}^*(\mathcal{R}\mathcal{R}^*)^{-1}\mathcal{A}_G\mathcal{B}\eta^*\|_{\mathcal{H}} \\
&\le \|\mathcal{I}_{G^\perp}\|_{\mathcal{H}\to L_2}\|\mathcal{R}^*(\mathcal{R}\mathcal{R}^*)^{-1}\|_{\ell_2\to\mathcal{H}}\|\mathcal{A}_G\|_{L_2\to\ell_2}\|\mathcal{B}\eta^*\|_{L_2} \\
&\lesssim \sqrt{\lambda_{p+1}}\frac{1}{\sqrt{\alpha_L}}\sqrt{n}b = \sqrt{\frac{n\lambda_{p+1}}{\alpha_L}}b.
\end{aligned}
$$

In the second-to-last-line, we used the result of Lemma A.5 to bound $\|\mathcal{B}\eta^*\|_{L_2}$. ∎

Using a sub-Gaussian tail bound, term B is thus bounded by $\sqrt{\frac{\lambda_{p+1}n\log n}{\alpha_L}}$ with probability at least $1 - \frac{1}{n^2}$. We then apply this to each of the n leave-one-out sets to obtain the bound on term B with probability $1 - \frac{1}{n}$.

For term C, we note, as in the proof of Lemma 4.4, that

$$|\mathcal{P}_G\hat{\eta}_0(x) - \overline{\mathcal{S}}\eta^*(x)| = |((\mathcal{B} - \overline{\mathcal{B}})\eta^*)(x)|,$$

where $\mathcal{B} := (\mathcal{I}_G + \mathcal{A}_G^*(\mathcal{R}\mathcal{R}^*)^{-1}\mathcal{A}_G)^{-1}$ (this follows from the definition of $\hat{\eta}$ and the push-through identity). This term hence has sub-Gaussian norm at most

$$\|(\mathcal{B} - \overline{\mathcal{B}})\eta^*\|_{L_2} \le \frac{bc}{1-c} \lesssim bc,$$

where we have used the result of Lemma A.5. So, using a tail-bound, we have that

$$|\mathcal{P}_G\hat{\eta}_0(x) - \overline{\mathcal{S}}\eta^*(x)| \lesssim bc\sqrt{\log n}$$

with probability at least $1 - \frac{1}{n^2}$ (so this holds for all the leave-one-out sets with probability $1 - \frac{1}{n}$). Combining the above bounds yields Theorem 3.4.

5. SVP with general regression functions. In the above results, we have assumed that $\eta^* \in G$ to facilitate clean analysis (and we note that some prior works on harmless interpolation in regression and classification make such assumptions in either an exact or approximate sense [33, 39]). However, this assumption can be relaxed in a straightforward way, resulting in an additional term in our main theorems. We consider in particular $\eta^* \in \mathcal{H}$ (so, more explicitly, $\|\eta^*\|_{\mathcal{H}} < \infty$). This bounded-RKHS-norm assumption is commonplace in classical kernel ridge regression analysis. In this case, we can decompose the MNI into the part that fits $\mathcal{P}_G\eta^*$ and ξ and the part that fits $\mathcal{P}_{G^\perp}\eta^*$ as follows:

$$\hat{\eta} = \underbrace{\mathcal{A}^*(\mathcal{A}\mathcal{A}^*)^{-1}\mathcal{A}_G\mathcal{P}_G\eta^* + \mathcal{A}^*(\mathcal{A}\mathcal{A}^*)^{-1}\xi}_{:=\hat{\eta}_{G,\xi}} + \underbrace{\mathcal{A}^*(\mathcal{A}\mathcal{A}^*)^{-1}\mathcal{R}\mathcal{P}_{G^\perp}\eta^*}_{:=\hat{\eta}_{G^\perp}}.$$

Now, based on Proposition 4.1, a sufficient condition for SVP is that for all i,

$$|(\hat{\eta}_{G,\xi})_{\backslash i}(x_i)| + |(\hat{\eta}_{G^\perp})_{\backslash i}(x_i)| \ll 1.$$

Our main theorems give conditions under which the first of these terms is small. So, to extend to general $\eta^* \in \mathcal{H}$, it suffices to give a condition for the second term, i.e., the contribution of the component of η^* orthogonal to G on $\hat{\eta}$, to be much less than 1. This is done in the following lemma.

Lemma 5.1. *Let $x \sim \mu$ be independent of the training data. Then, the following bounds hold for the MNI:*

1. **BOS assumption:**

$$|\hat{\eta}_{G^\perp}(x)| \lesssim \left(\sqrt{\frac{\alpha_U p}{n}} + \frac{1}{\sqrt{\tilde{\alpha}}} \|\mathcal{R} k_x^R\|_2 \right) \|\mathcal{P}_{G^\perp} \eta^*\|_{\mathcal{H}}.$$

2. **SG assumption:** *With probability at least $1 - \frac{1}{n^2}$,*

$$|\hat{\eta}_{G^\perp}(x)| \lesssim \left(\frac{\lambda_1 \sqrt{n\tilde{\alpha}}}{\tilde{\alpha} + n\lambda_1} + \sqrt{\lambda_{p+1}} \right) \|\mathcal{P}_{G^\perp} \eta^*\|_{\mathcal{H}} \sqrt{\log n},$$

where $\tilde{\alpha} = \frac{\alpha_L + \alpha_U}{2}$ and $k_x^R = \mathcal{P}_{G^\perp} k_x$.

Proof.

1. **BOS:** Observe that

$$\begin{aligned}
|\hat{\eta}_{G^\perp}(x)| &= |\langle \mathcal{A}^* (\mathcal{A}\mathcal{A}^*)^{-1} \mathcal{R} \mathcal{P}_{G^\perp} \eta^*, k_x \rangle_{\mathcal{H}}| \\
&= |\langle \mathcal{R} \mathcal{P}_{G^\perp} \eta^*, (\mathcal{A}\mathcal{A}^*)^{-1} \mathcal{A} k_x \rangle_{\ell_2}| \\
&\leq \|\mathcal{R} \mathcal{P}_{G^\perp} \eta^*\|_2 \|(\mathcal{A}\mathcal{A}^*)^{-1} \mathcal{A} k_x\|_2 \\
&\leq \|\mathcal{R}\|_{\mathcal{H} \to \ell_2} \|\mathcal{P}_{G^\perp} \eta^*\|_{\mathcal{H}} \|(\mathcal{A}\mathcal{A}^*)^{-1} \mathcal{A} k_x\|_2 \\
&\lesssim \sqrt{\alpha_U} \|\mathcal{P}_{G^\perp} \eta^*\|_{\mathcal{H}} \left(\sqrt{\frac{p}{n}} + \frac{1}{\tilde{\alpha}} \|\mathcal{R} k_x^R\|_2 \right),
\end{aligned}$$

   where the bound on $\|(\mathcal{A}\mathcal{A}^*)^{-1} \mathcal{A} k_x\|_2$ in the last inequality is derived in the proof of Lemma 4.2.

2. **SG:** Since k_x is a zero-mean sub-Gaussian, with probability at least $1 - \frac{1}{n^2}$,

$$\begin{aligned}
|\hat{\eta}_{G^\perp}(x)| &\lesssim \|\mathcal{A}^* (\mathcal{A}\mathcal{A}^*)^{-1} \mathcal{R} \mathcal{P}_{G^\perp} \eta^*\|_{L_2} \sqrt{\log n} \\
&= \|\mathcal{A}^* (\tilde{\alpha} I_n + \mathcal{A}_G \mathcal{A}_{G*})^{-1} (\tilde{\alpha} I_n + \mathcal{A}_G \mathcal{A}_{G*}) (\mathcal{A}\mathcal{A}^*)^{-1} \mathcal{R} \mathcal{P}_{G^\perp} \eta^*\|_{L_2} \sqrt{\log n} \\
&\lesssim \|\mathcal{A}^* (\tilde{\alpha} I_n + \mathcal{A}_G \mathcal{A}_{G*})^{-1}\|_{\ell_2 \to L_2} \|\mathcal{R}\|_{\mathcal{H} \to \ell_2} \|\mathcal{P}_{G^\perp} \eta^*\|_{\mathcal{H}} \sqrt{\log n},
\end{aligned}$$

   where the last inequality follows from Lemma 8 of [31]. Now,

$$\begin{aligned}
\|\mathcal{A}_G^* (\tilde{\alpha} I_n + \mathcal{A}_G \mathcal{A}_{G*})^{-1}\|_{\ell_2 \to L_2} &= \|(\tilde{\alpha} I_G + \mathcal{A}_G * \mathcal{A}_G)^{-1} \mathcal{A}_G^*\|_{\ell_2 \to L_2} \\
&= \|(\tilde{\alpha} \mathcal{T}_G^{-1} + \mathcal{C}^* \mathcal{C})^{-1} \mathcal{C}^*\|_{\ell_2 \to L_2} \\
&\lesssim \frac{\sqrt{n}}{\tilde{\alpha} \lambda_1^{-1} + n},
\end{aligned}$$

by part 3 of Assumption 3.1. Also,

$$\|\mathcal{R}^*(\tilde{\alpha}I_n + \mathcal{A}_G\mathcal{A}_{G^*})^{-1}\|_{\ell_2 \to L_2} \leq \frac{1}{\tilde{\alpha}}\|\mathcal{R}\|_{\ell_2 \to L_2} \lesssim \sqrt{\frac{\lambda_{p+1}}{\tilde{\alpha}}}.$$

Combining the above yields, with probability at least $1 - \frac{1}{n^2}$,

$$|\langle \mathcal{A}^*(\mathcal{A}\mathcal{A}^*)^{-1}\mathcal{R}\mathcal{P}_{G^\perp}\eta^*, k_x\rangle_{\mathcal{H}}| \lesssim \left(\frac{\sqrt{n}}{\tilde{\alpha}\lambda_1^{-1} + n} + \sqrt{\frac{\lambda_{p+1}}{\tilde{\alpha}}}\right)\sqrt{\tilde{\alpha}}\|\mathcal{P}_{G^\perp}\eta^*\|_{\mathcal{H}}\sqrt{\log n}.$$ ∎

Applying the above lemma to each of the leave-one-out data sets and recalling that $\|\mathcal{R}_{\backslash i}k_{x_i}^R\|_2 \leq \alpha_U - \alpha_L$, we obtain the following additional sufficient condition for SVP:

1. **BOS:**

$$\left(\sqrt{\frac{\alpha_U p}{n}} + \frac{\alpha_U - \alpha_L}{\sqrt{\tilde{\alpha}}}\right)\|\mathcal{P}_{G^\perp}\eta^*\|_{\mathcal{H}} \ll 1.$$

2. **SG:**

$$\left(\frac{\lambda_1\sqrt{n\tilde{\alpha}}}{\tilde{\alpha} + n\lambda_1} + \sqrt{\lambda_{p+1}}\right)\|\mathcal{P}_{G^\perp}\eta^*\|_{\mathcal{H}}\sqrt{\log n} \ll 1.$$

Note that this condition is quite mild in the sub-Gaussian case, and it is always satisfied for large enough n under the bilevel ensemble (with the exception of the isotropic case $q = \beta - r$, for which the covariance is no longer bilevel). So, in this case we still have SVP as long as there exists an $r < 1$ such that $\mathcal{P}_G\eta^*$ satisfies Assumption 3.1. This is consistent with the results of earlier works, which do not assume any label-generating process for sub-Gaussian data. In the BOS case, this condition is somewhat more restrictive but still allows for nontrivial regimes in which SVP can occur for $\eta^* \notin G$, e.g., in the bilevel ensemble when $(1-r)/2 < q < 1 - 2r$ (implicitly needing $r < 1/3$) and $\beta > 2 + r + q$.

6. Simulations and discussion. To verify our theoretical findings, we now perform simulations of the MNI and SVM procedures under the bilevel model. We consider the case where the input space X is the interval $[0,1]$ with the uniform measure. Furthermore, we consider the Fourier featurization, specified by the eigenfunctions[6] $v_\ell(x) = e^{j2\pi\ell x}$ for $\ell = -d, \ldots, d$. Performing MNI or SVM with these features and the bilevel weighting on features corresponds to the kernel function

$$k(x,y) = \sum_{\ell=-d}^{d} \lambda_\ell v_\ell(x)\overline{v_\ell(y)} = \sum_{\ell=-d}^{d} \lambda_\ell e^{j2\pi\ell(x-y)},$$

where the λ_ℓ are given as in (3.2) for $\ell > 0$ and $\lambda_{-\ell} = \lambda_\ell$. For all experiments, we use a training size of $n = 100$ (note that the choice of n affects the kernel eigenvalues).

---

[6]For simplicity, we use complex exponential eigenfunctions, but we will always use complex-conjugate–symmetric coefficients, resulting in real functions. We could use purely real sines and cosines as our eigenfunctions at the expense of somewhat more complicated notation.

We first generate a random ground-truth function η^* as a linear combination of $\{v_\ell\}_{\ell=-3}^3$, with standard normal coefficients, scaled so that $\eta^*(x) \in [-1,1]$. For each x_i in the training set, we generate y_i according to the rule $\mathbf{P}(y_i = 1) = \frac{1+\eta^*(x_i)}{2}$. We then compute and plot the MNI and SVM estimates for three choices of bilevel parameters, $(\beta, r, q) = (3.2, 0.4, -0.4)$, $(3.2, 0.4, 0.4)$, and $(3.2, 0.4, 0.8)$, corresponding to different levels of effective overparameterization. The results are shown in Figure 1. Based on Corollary 3.3, the two estimates should coincide with high probability when $q > \frac{1-0.4}{2} = 0.3$. Indeed, from Figures 1b and 1c, we see that this is the case for the second and third choices of parameters. Our results do not draw any conclusions about the occurrence of SVP in the range $q < 0.3$, but we see that in this example, SVP does not occur for a very low degree of effective overparameterization, as shown in Figure 1a.

To measure the effect of the bilevel ensemble parameters, we also plot a heatmap of SVP occurrence for various values of r (the number of favored features) and q (the effective overparameterization). Here, we fix $n = 100$, $\beta = 3.2$, and the ground-truth function η^* (generated as in Figure 1), and we plot the proportion of trials (out of 25) in which SVP occurs for each (r, q). A heatmap of these results is provided in Figure 2 along with an overlay
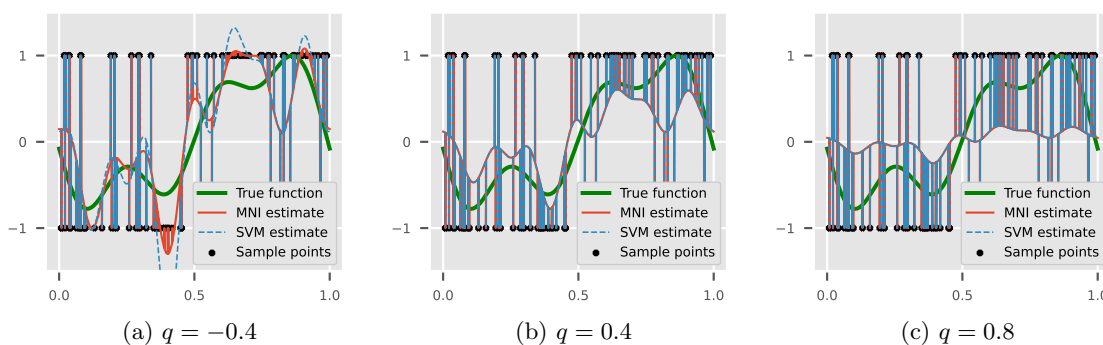

(a) $q = -0.4$ (b) $q = 0.4$ (c) $q = 0.8$

Figure 1. *SVM and MNI solutions with Fourier features in the bilevel ensemble for various levels of effective overparameterization.*
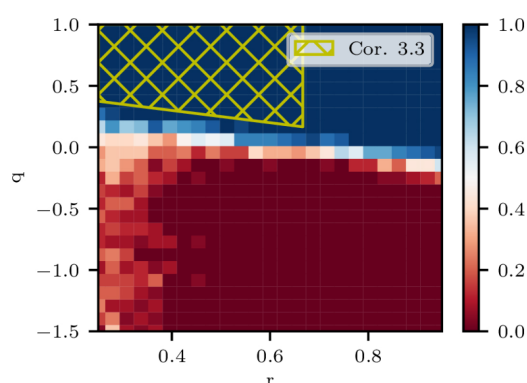

Figure 2. *Heatmap showing the proportion of trials (out of 25) in which SVP occurs for varying parameter choices in the Fourier bilevel ensemble, and overlay of the SVP regime predicted by our results.*

of the regime for which Corollary 3.3 predicts SVP. We note that, at least for the case of Fourier features, there exist values of (r,q) for which SVP occurs but which are not predicted by our theoretical results; understanding SVP in these regimes, e.g., $\frac{2}{3} < r < 1$ (and, more generally, providing sharp lower bounds) is an interesting and important direction for future work.

Appendix A. Concentration bounds and technical lemmas. In this appendix we state some useful concentration bounds which can be used to verify the operator concentration conditions in Assumption 3.1. These inequalities and their proofs appear in [31]. Similar concentration bounds for sub-Gaussian/independent features can also be found in [4, 39].

Lemma A.1 (general residual Gram matrix concentration). *Let $k^R(x,y)$ be the restriction of the kernel k to $G^\perp$. Then,*

$$\mathbf{E}\|\mathcal{R}\mathcal{R}^* - (\mathrm{tr}\mathcal{T}_{G^\perp})I_n\|^2 \lesssim n^2\mathrm{tr}(\mathcal{T}_{G^\perp}^2) + \|k^R(\cdot,\cdot) - \mathrm{tr}\mathcal{T}_{G^\perp}\|_\infty^2,$$

where $\|k^R(\cdot,\cdot) - \mathrm{tr}\mathcal{T}_{G^\perp}\|_\infty = \sup_x\{|k^R(x,x) - \mathrm{tr}\,\mathcal{T}_{G^\perp}|\}$.

Lemma A.2 (residual Gram matrix concentration for independent, sub-Gaussian features). *Suppose the features $\{v_\ell(x)\}_{\ell\geq 1}$ are zero-mean, independent, and sub-Gaussian. Then, for $t>0$, with probability at least $1 - e^{-t}$,*

$$\|\mathcal{R}\mathcal{R}^* - (\mathrm{tr}\mathcal{T}_{G^\perp})I_n\| \lesssim \sqrt{(n+t)\mathrm{tr}(\mathcal{T}_{G^\perp}^2)} + (n+t)\lambda_{p+1}.$$

Lemma A.3 (sampling operator concentration on G in a BOS). *Suppose the kernel features $\{v_\ell(x)\}_{\ell\geq 1}$ satisfy the BOS condition with constant $C\geq 1$. Then, for $t>0$, with probability at least $1 - e^{-t}$,*

$$\frac{1}{n}\|\mathcal{C}^*\mathcal{C} - n\mathcal{I}_G\|_{L_2} \lesssim \sqrt{\frac{Cp(t+\log p)}{n}} + \frac{Cp(t+\log p)}{n}.$$

Lemma A.4 (sampling operator concentration on G for independent, sub-Gaussian features). *Suppose the features $\{v_\ell(x)\}_{\ell\geq 1}$ are zero-mean, independent, and sub-Gaussian. Then, with probability at least $1 - e^{-t}$,*

$$\frac{1}{n}\|\mathcal{C}^*\mathcal{C} - n\mathcal{I}_G\|_{L_2} \lesssim \sqrt{\frac{p+t}{n}} + \frac{p+t}{n}.$$

We next state the following approximation result, proven in [31], which we use frequently to prove various technical lemmas.

Lemma A.5 (Lemma 7 in [31]). *Let $\mathcal{B} := (\mathcal{I}_G + \mathcal{A}_G^*(\mathcal{R}\mathcal{R}^*)^{-1}\mathcal{A}_G)^{-1}$ and $\overline{\mathcal{B}} = (\mathcal{I}_G + \frac{n}{\alpha}\mathcal{T}_G)^{-1}$. Then, under the conditions of Assumption 3.1,*

$$\mathcal{B} - \overline{\mathcal{B}} = \left(\sum_{k=1}^\infty (-\mathcal{Q})^k\right)\overline{\mathcal{B}}$$

for $\mathcal{Q} := (\mathcal{T}_G^{-1} + \frac{n}{\alpha}\mathcal{I}_G)^{-1}(\mathcal{C}^*(\mathcal{R}\mathcal{R}^*)^{-1}\mathcal{C} - \frac{n}{\alpha}\mathcal{I}_G)$ satisfying $\|\mathcal{Q}\|_{L_2} \leq c$. Here, $c < 1$ is an upper bound on the quantity $\frac{\alpha_U - \alpha_L}{\alpha_U + \alpha_L} + \frac{2}{n}\|\mathcal{C}^*\mathcal{C} - n\mathcal{I}_G\|_{L_2}$.

Note that by applying the formula for the sum of an infinite geometric series, the above result implies that $\|(\mathcal{B} - \overline{\mathcal{B}})\eta^*\|_{L_2} \leq \frac{c}{1-c}\|\overline{\mathcal{B}}\eta^*\|_{L_2}$ and $\|\mathcal{B}\eta^*\|_{L_2} \leq (1 + \frac{c}{1-c})\|\overline{\mathcal{B}}\eta^*\|_{L_2}$.

Appendix B. Proofs for the bilevel ensemble.

Proof of Corollary 3.3. We bound each term for the bilevel ensemble as in [31, Corollary 1]. First, note that

$$n\sqrt{\operatorname{tr}(\mathcal{T}_{G^{\perp}}^2)} \asymp n \cdot ((n^{\beta} - n^r) \cdot n^{-2\beta+2r+2q})^{1/2} \asymp n^{1-\frac{\beta}{2}+r+q}.$$

By the thin-shell assumption that $\sup_x \left| \sum_{\ell>p}(v_{\ell}(x)^2 - 1) \right| \lesssim n^{\beta/2+1}$,

$$\sup_x |k^R(x,x) - \operatorname{tr}\mathcal{T}_{G^{\perp}}| = \sup_x \left| \sum_{\ell>p} \lambda_{\ell}(v_{\ell}(x)^2 - 1) \right| = n^{-\beta+r+q} \sup_x \left| \sum_{\ell>p}(v_{\ell}(x)^2 - 1) \right|$$
$$\lesssim n^{-\beta+r+q} n^{\beta/2+1} = n^{1-\beta/2+r+q},$$

so Lemma A.1 implies that in the limit as $n \to \infty$, we can choose $\alpha_L = \lambda_{min}(\mathcal{R}\mathcal{R}^*)$ and $\alpha_U = \lambda_{max}(\mathcal{R}\mathcal{R}^*)$ with $\alpha_U, \alpha_L, \bar{\alpha} \asymp n^{r+q}$ as long as $1 - \frac{\beta}{2} + r + q < r + q$, which is true by our assumption that $\beta > 2$. More precisely, Lemma A.1 gives us an expectation bound, which we can then use with Markov's inequality to get a high-probability bound. We can obtain the stated bounds within factors of n^{ϵ} for arbitrarily small $\epsilon > 0$, and the probability of these bounds holding goes to 1 as $n \to \infty$. We can now bound the term c from Assumption 3.1. Lemma A.1 implies that

$$\frac{\alpha_U - \alpha_L}{\alpha_U + \alpha_L} \lesssim \frac{1}{\bar{\alpha}} n\sqrt{\operatorname{tr}(\mathcal{T}_{G^{\perp}}^2)} \asymp n^{1-\frac{\beta}{2}}.$$

Then, applying Lemma A.3 for each of the n leave-one-out sets allows us to bound c as

$$c \lesssim \frac{\alpha_U - \alpha_L}{\alpha_U + \alpha_L} + \frac{2}{n}\|\mathcal{C}_{\setminus i}^* \mathcal{C}_{\setminus i} - n\mathcal{I}_G\|_{L_2} \lesssim n^{1-\frac{\beta}{2}} + n^{\frac{r-1}{2}}\sqrt{\log n}.$$

Now, note that for the bilevel ensemble, we have $b = \|\frac{\bar{\alpha}}{\bar{\alpha}+n}\mathcal{I}_G \eta^*\|_{L_{\infty}} = \frac{\bar{\alpha}}{\bar{\alpha}+n}\|\eta^*\|_{L_{\infty}} \asymp \frac{\bar{\alpha}}{\bar{\alpha}+n} \asymp \min\{1, n^{r+q-1}\}$. We consider the following two cases (ignoring log factors, which are negligible compared to powers of n as $n \to \infty$):

1. If $q > 1 - r$, $b \asymp 1$, so the conditions for SVP from Theorem 3.2 can be written as

$$n^{(r-1)/2} + n^{1-\beta/2}n^{1/2} + (n^{2-\beta} + n^{r-1})n^{r/2} \ll 1.$$

One can easily check that this holds as $n \to \infty$ when $r < 2/3$ and $\beta > 3$.

2. If $q \leq 1 - r$, then $b \asymp n^{r+q-1}$, so the condition is

$$n^{1-r-q}n^{(r-1)/2} + n^{1-\beta/2}(n^{1-r-q} + n^{1/2}) + (n^{2-\beta} + n^{r-1})n^{r/2} \ll 1.$$

This holds as $n \to \infty$ for $r < 2/3$, $q > \frac{1-r}{2}$, and $\beta > 3$.

Combining the above two cases yields the result. ∎

Proof of Corollary 3.5. The proof of this statement is similar to that of Corollary 3.3, but it uses sharper concentration bounds from Lemmas A.2 and A.4 for independent, sub-Gaussian features. Again, we first bound the quantities $\alpha_U - \alpha_L$ and c. Lemma A.2 implies that with probability at least $1 - n^{-1}$,

$$\|\mathcal{R}\mathcal{R}^* - (\operatorname{tr}\mathcal{T}_{G^{\perp}})I_n\| \lesssim n^{1/2-\beta/2+r+q} + n^{1-\beta+r+q}.$$

Since $\beta > 1$, we can again choose $\alpha_L = \lambda_{min}(\mathcal{R}\mathcal{R}^*)$ and $\alpha_U = \lambda_{max}(\mathcal{R}\mathcal{R}^*)$ with $\alpha_U, \alpha_L, \overline{\alpha} \asymp n^{r+q}$. Furthermore,

$$\frac{\alpha_U - \alpha_L}{\alpha_U + \alpha_L} \lesssim \frac{1}{n^{r+q}} n^{(1-\beta)/2+r+q} \asymp n^{(1-\beta)/2},$$

and Lemma A.4 (again applied as a union bound over the n leave-one-out sets) implies that we can take $c \asymp n^{(1-\beta)/2} + n^{\frac{r-1}{2}} \gtrsim \frac{\alpha_U - \alpha_L}{\alpha_U + \alpha_L} + \frac{2}{n} \|\mathcal{C}_{\backslash i}^* \mathcal{C}_{\backslash i} - n\mathcal{I}_G\|_{L_2}$. We again consider the following two cases from the previous corollary (again omitting log factors):

1. If $q > 1 - r$, $b \asymp 1$, the conditions for SVP from Theorem 3.4 can be written as

$$n^{(r-1)/2} + n^{(1-\beta)/2} + \sqrt{\frac{n^{-\beta+r+q+1}}{n^{r+q}}} + (n^{(1-\beta)/2} + n^{\frac{r-1}{2}}) \ll 1.$$

Here, the left-hand side is always decreasing in n for $\beta > 1$ and $r < 1$.

2. If $q \le 1 - r$, then $b \asymp n^{r+q-1}$, so the condition is

$$n^{\frac{r-1}{2}+1-r-q} + n^{1-r-q} \cdot n^{(1-\beta)/2} + \sqrt{\frac{n^{-\beta+r+q+1}}{n^{r+q}}} + (n^{(1-\beta)/2} + n^{\frac{r-1}{2}}) \ll 1.$$

In this case, the left-hand side is decreasing for $q > \frac{1-r}{2}$, and $\beta > 3 - 2r - 2q$. Combining the above two cases, we arrive at the desired result. ∎

Proof of Corollary 3.6. To prove this corollary, we first characterize the range of parameters (n, β, q, r) for which the excess classification risk of the MNI solution goes to 0 in the limit as $n \to \infty$. In the BOS case, it was proven in Corollary 1 of [31] that this occurs in the following two settings (the full finite-sample bound for the MNI is given in [31]):

1. $q < 1 - r$, $\beta > 2$ (for which regression is also consistent).
2. $1 - r < q < \frac{3}{2}(1-r)$ and $\beta > 2(r+q)$ (for which classification is consistent but regression is not).

Combining these with the conditions for SVP in bounded orthonormal systems from Corollary 3.3, we can immediately conclude that the SVM solution is classification-consistent in the regime where both conditions hold (i.e., when SVP occurs and when the MNI is consistent). Part 1 of Corollary 3.6 follows.

Now, to prove part 2 of the corollary, we modify the proof for the BOS case by using sharper concentration bounds for independent, sub-Gaussian features. We first recall the following expression of the excess risk from [31]: $\mathcal{E} \le \frac{\|\hat{\eta}_r\|_{L_2}}{s}$, where $s > 0$ is a parameter we can choose, and $\hat{\eta} = s\eta^* + \hat{\eta}_r$.

Note that the idealized survival operator in the bilevel ensemble is given by $\overline{\mathcal{S}} = \frac{n}{n+\overline{\alpha}} \mathcal{I}_G \asymp \frac{n}{n+n^{r+q}} \mathcal{I}_g \asymp \min(1, n^{1-r-q})\mathcal{I}_G$. So, we will apply the above bound on the excess classification risk with the choice $s = \min(1, n^{1-r-q})$. Then, by definition, $\hat{\eta}_r = \epsilon + \hat{\eta}_0 - \overline{\mathcal{S}}\eta^*$, so

$$\|\hat{\eta}_r\|_{L_2} \le \|\epsilon\|_{L_2} + \|\hat{\eta}_0 - \overline{\mathcal{S}}\eta^*\|_{L_2}.$$

The first term can be bounded by Theorem 2 of [31], which yields

$$\|\epsilon\|_{L_2} \lesssim \sqrt{\frac{p}{n}} + \left(\frac{n \sum_{\ell > p} \lambda_\ell^2}{\left(\sum_{\ell > p} \lambda_\ell \right)^2} \right)^{1/2} \asymp n^{(r-1)/2} + \left(\frac{n n^\beta n^{-2\beta+2r+2q}}{n^{2\beta} n^{-2\beta+2r+2q}} \right)^{1/2} = n^{(r-1)/2} + n^{(1-\beta)/2},$$

where we have used the scaling derived in the proof of Corollary 3.5.

Furthermore, we can bound $\|\hat{\eta}_0 - \overline{\mathcal{S}}\eta^*\|_{L_2}$ with Lemma 6 of [31] as follows:

$$
\begin{aligned}
\|\hat{\eta}_0 - \overline{\mathcal{S}}\eta^*\|_{L_2} &\lesssim \left(c + \sqrt{\frac{n\lambda_{p+1}}{\overline{\alpha}}} \right) \min\left(\lambda_1, \frac{\overline{\alpha}}{n\sqrt{\lambda_p}}, \sqrt{\frac{\overline{\alpha}}{n}} \right) \|\eta^*\|_{\mathcal{H}} \\
&\lesssim \left(n^{(1-\beta)/2} + n^{\frac{r-1}{2}} + n^{(1-\beta)/2} \right) \min\left(1, n^{r+q-1}, n^{(r+q-1)/2} \right) \|\eta^*\|_{L_2} \\
&\lesssim \left(n^{(1-\beta)/2} + n^{\frac{r-1}{2}} \right) \min\left(1, n^{r+q-1} \right) \|\eta^*\|_{L_2}.
\end{aligned}
$$

So, we can obtain the risk bound

$$
\mathcal{E} \lesssim \max(1, n^{r+q-1}) \left(n^{(r-1)/2} + n^{(1-\beta)/2} + \left(n^{(1-\beta)/2} + n^{\frac{r-1}{2}} \right) \min\left(1, n^{r+q-1} \right) \right) \|\eta^*\|_{L_2}.
$$

From this we obtain the following:
1. If $q < 1 - r$, then $\mathcal{E} \to 0$ (and, in fact, $s \to 1$).
2. if $q > 1 - r$, then $\mathcal{E} \to 0$ if $q < \frac{3}{2}(1 - r)$ and $q < 1 + \frac{\beta-1}{2} - r$.

Combining this range of allowable parameters with the conditions for SVP in Corollary 3.5 yields the part 2 of the corollary. Again, the case of $q \leq \frac{1-r}{2} < 1 - r$ is covered by previous results as discussed in [21]. ∎

REFERENCES

[1] N. ARDESHIR, C. SANFORD, AND D. J. HSU, *Support vector machines and linear regression coincide with very high-dimensional features*, in Advances in Neural Information Processing Systems 34 (NeurIPS 2021), M. Ranzato et al., eds., Curran Associates, 2021, pp. 4907–4918.

[2] J.-Y. AUDIBERT AND A. B. TSYBAKOV, *Fast learning rates for plug-in classifiers*, Ann. Statist., 35 (2007), pp. 608–633, https://doi.org/10.1214/009053606000001217.

[3] D. BARTL AND S. MENDELSON, *Random embeddings with an almost Gaussian distortion*, Adv. Math., 400 (2022), 108261, https://doi.org/10.1016/j.aim.2022.108261.

[4] P. BARTLETT, P. LONG, G. LUGOSI, AND A. TSIGLER, *Benign overfitting in linear regression*, Proc. Natl. Acad. Sci. USA, 117 (2020), pp. 30063–30070, https://doi.org/10.1073/pnas.1907378117.

[5] P. BARTLETT AND J. SHAWE-TAYLOR, *Generalization performance of support vector machines and other pattern classifiers*, in Advances in Kernel Methods: Support Vector Learning, B. Schölkopf and C. J. C. Burges, eds., MIT Press, 1999, pp. 43–54.

[6] M. BELKIN, *Fit without fear: Remarkable mathematical phenomena of deep learning through the prism of interpolation*, Acta Numer., 30 (2021), pp. 203–248, https://doi.org/10.1017/s0962492921000039.

[7] M. BELKIN, D. HSU, AND J. XU, *Two models of double descent for weak features*, SIAM J. Math. Data Sci., 2 (2020), pp. 1167–1180, https://doi.org/10.1137/20M1336072.

[8] M. BELKIN, S. MA, AND S. MANDAL, *To understand deep learning we need to understand kernel learning*, in Proceedings of the 35th International Conference on Machine Learning (ICML), PMLR 80, 2018, pp. 541–549.

[9] A. BUHOT AND M. B. GORDON, *Robust learning and generalization with support vector machines*, J. Phys. A Math. Gen., 34 (2001), pp. 4377–4388.

[10] Y. CAO, Q. GU, AND M. BELKIN, *Risk bounds for over-parameterized maximum margin classification on sub-Gaussian mixtures*, in Advances in Neural Information Processing Systems 34 (NeurIPS 2021), M. Ranzato et al., eds., Curran Associates, 2021, pp. 8407–8418, https://proceedings.neurips.cc/paper/2021/hash/46e0eae7d5217c79c3ef6b4c212b8c6f-Abstract.html.

[11] N. S. CHATTERJI AND P. M. LONG, *Finite-sample analysis of interpolating linear classifiers in the over-parameterized regime*, J. Mach. Learn. Res., 22 (2021), pp. 1–30.

[12] C. CORTES AND V. VAPNIK, *Support-vector networks*, Mach. Learn., 20 (1995), pp. 273–297.

[13] Y. DAR, V. MUTHUKUMAR, AND R. G. BARANIUK, *A Farewell to the Bias-Variance Tradeoff? An Overview of the Theory of Overparameterized Machine Learning*, preprint, https://arxiv.org/abs/2109.02355, 2021.

[14] R. DIETRICH, M. OPPER, AND H. SOMPOLINSKY, *Statistical mechanics of support vector networks*, Phys. Rev. Lett., 82 (1999), pp. 2975–2978.

[15] P. GERMAIN, A. LACOSTE, F. LAVIOLETTE, M. MARCHAND, AND S. SHANIAN, *A PAC-Bayes sample-compression approach to kernel methods*, in Proceedings of the 28th International Conference on Machine Learning, Bellevue, WA, 2011, pp. 297–304.

[16] P. GOLIK, P. DOETSCH, AND H. NEY, *Cross-entropy vs. squared error training: A theoretical and experimental comparison*, in Proceedings of Interspeech, Lyon, France, 2013, pp. 1756–1760.

[17] T. GRAEPEL, R. HERBRICH, AND J. SHAWE-TAYLOR, *PAC-Bayesian compression bounds on the prediction error of learning algorithms for classification*, Mach. Learn., 59 (2005), pp. 55–76.

[18] S. GUNASEKAR, J. D. LEE, D. SOUDRY, AND N. SREBRO, *Implicit bias of gradient descent on linear convolutional networks*, in Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada, Adv. Neural Inf. Process. Syst. 31, S. Bengio et al., eds., Curran Associates, 2018, pp 9461–9471.

[19] T. HASTIE, A. MONTANARI, S. ROSSET, AND R. J. TIBSHIRANI, *Surprises in high-dimensional ridgeless least squares interpolation*, Ann. Statist., 50 (2022), pp. 949–986, https://doi.org/10.1214/21-AOS2133.

[20] D. HSU, S. M. KAKADE, AND T. ZHANG, *Random design analysis of ridge regression*, Found. Comput. Math., 14 (2014), pp. 569–600, https://doi.org/10.1007/s10208-014-9192-1.

[21] D. HSU, V. MUTHUKUMAR, AND J. XU, *On the proliferation of support vectors in high dimensions*, in Proceedings of The 24th International Conference on Artificial Intelligence and Statistics (AISTATS), PMLR 130, 2021, pp. 91–99, http://proceedings.mlr.press/v130/hsu21a.html.

[22] L. HUI AND M. BELKIN, *Evaluation of neural architectures trained with square loss vs. cross-entropy in classification tasks*, in Proceedings of the International Conference on Learning Representations (ICLR), 2021, https://openreview.net/forum?id=hsFN92eQEla.

[23] Z. JI, M. DUDÍK, R. E. SCHAPIRE, AND M. TELGARSKY, *Gradient descent follows the regularization path for general losses*, in Proceedings of the Thirty Third Conference on Learning Theory, PMLR 125, 2020, pp. 2109–2136.

[24] Z. JI AND M. TELGARSKY, *The implicit bias of gradient descent on nonseparable data*, in Proceedings of the Thirty-Second Conference on Learning Theory, PMLR 99, 2019, pp. 1772–1798.

[25] Z. JI AND M. TELGARSKY, *Directional convergence and alignment in deep learning*, in Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada, Adv. Neural Inf. Process. Syst. 33, H. Larochelle et al., eds., Curran Associates, 2020, pp. 17176–17186.

[26] D. M. KLINE AND V. L. BERARDI, *Revisiting squared-error and cross-entropy functions for training neural network classifiers*, Neural Comput. Appl., 14 (2005), pp. 310–318.

[27] K.-W. LAI AND V. MUTHUKUMAR, *General Loss Functions Lead to (Approximate) Interpolation in High Dimensions*, preprint, arXiv:2303.07475, 2023.

[28] T. LIANG AND A. RAKHLIN, *Just interpolate: Kernel "ridgeless" regression can generalize*, Ann. Statist., 48 (2020), pp. 1329–1347, https://doi.org/10.1214/19-aos1849.

[29] D. MALZAHN AND M. OPPER, *A statistical physics approach for the analysis of machine learning algorithms on real data*, J. Stat. Mech. Theory Exp., 2005 (2005), P11001.

[30] D. MCALLESTER, *Simplified PAC-Bayesian margin bounds*, in Learning Theory and Kernel Machines, Lect. Notes Comput. Sci. 2777, Springer, 2003, pp. 203–215.

[31] A. D. MCRAE, S. KARNIK, M. A. DAVENPORT, AND V. MUTHUKUMAR, *Harmless interpolation in regression and classification with structured features*, in Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS), PMLR 151, 2022, pp. 5853–5875, https://proceedings.mlr.press/v151/mcrae22a.html.

[32] S. MEI, T. MISIAKIEWICZ, AND A. MONTANARI, *Generalization error of random feature and kernel methods: Hypercontractivity and kernel matrix concentration*, Appl. Comput. Harmon. Anal., 59 (2022), pp. 3–84, https://doi.org/10.1016/j.acha.2021.12.003.

[33] V. MUTHUKUMAR, A. NARANG, V. SUBRAMANIAN, M. BELKIN, D. HSU, AND A. SAHAI, *Classification vs. regression in overparameterized regimes: Does the loss function matter?*, J. Mach. Learn. Res., 22 (2021), pp. 1–69, https://jmlr.csail.mit.edu/papers/volume22/20-603/20-603.pdf.

[34] V. MUTHUKUMAR, K. VODRAHALLI, V. SUBRAMANIAN, AND A. SAHAI, *Harmless interpolation of noisy data in regression*, IEEE J. Sel. Areas Inf. Theory, 1 (2020), pp. 67–83, https://doi.org/10.1109/jsait.2020.2984716.

[35] B. NEYSHABUR, R. TOMIOKA, AND N. SREBRO, *In Search of the Real Inductive Bias: On the Role of Implicit Regularization in Deep Learning*, preprint, arXiv:1412.6614, 2014.

[36] R. M. RIFKIN, *Everything Old Is New Again: A Fresh Look at Historical Approaches in Machine Learning*, Ph.D. thesis, Massachusetts Institute of Technology, 2002.

[37] D. SOUDRY, E. HOFFER, M. S. NACSON, S. GUNASEKAR, AND N. SREBRO, *The implicit bias of gradient descent on separable data*, J. Mach. Learn. Res., 19 (2018), pp. 2822–2878.

[38] I. STEINWART AND C. SCOVEL, *Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs*, Constr. Approx., 35 (2012), pp. 363–417, https://doi.org/10.1007/s00365-012-9153-3.

[39] A. TSIGLER AND P. L. BARTLETT, *Benign overfitting in ridge regression*, J. Mach. Learn. Res., 24 (2023), pp. 1–76, https://www.jmlr.org/papers/volume24/22-1398/22-1398.pdf.

[40] R. VERSHYNIN, *High-Dimensional Probability*, Cambridge University Press, 2018.

[41] K. WANG, V. MUTHUKUMAR, AND C. THRAMPOULIDIS, *Benign overfitting in multiclass classification: All roads lead to interpolation*, in Advances in Neural Information Processing Systems 34 (NeurIPS 2021), M. Ranzato et al., eds., Curran Associates, 2021, pp. 24164–24179.

[42] K. WANG AND C. THRAMPOULIDIS, *Benign overfitting in binary classification of Gaussian mixtures*, in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 4030–4034, https://doi.org/10.1109/icassp39728.2021.9413946.

[43] C. ZHANG, S. BENGIO, M. HARDT, B. RECHT, AND O. VINYALS, *Understanding deep learning requires rethinking generalization*, in Proceedings of the International Conference on Learning Representations (ICLR), 2017, https://openreview.net/pdf?id=Sy8gdB9xx.

[44] T. ZHANG, *Learning bounds for kernel regression using effective data dimensionality*, Neural Comput., 17 (2005), pp. 2077–2098, https://doi.org/10.1162/0899766054323008.