Generalization for Multiclass Classification with Overparameterized Linear Models

Vignesh Subramanian
Department of Electrical Engineering and Computer Sciences
University of California Berkeley
Berkeley, CA-94720, USA
vignesh.subramanian@eecs.berkeley.edu

Rahul Arya
Department of Electrical Engineering and Computer Sciences
University of California Berkeley
Berkeley, CA-94720, USA
rahularya@berkeley.edu

Anant Sahai
Department of Electrical Engineering and Computer Sciences
University of California Berkeley
Berkeley, CA-94720, USA
sahai@eecs.berkeley.edu

Abstract

Via an overparameterized linear model with Gaussian features, we provide conditions for good generalization for multiclass classification of minimum-norm interpolating solutions in an asymptotic setting where both the number of underlying features and the number of classes scale with the number of training points. The survival/contamination analysis framework for understanding the behavior of overparameterized learning problems is adapted to this setting, revealing that multiclass classification qualitatively behaves like binary classification in that, as long as there are not too many classes (made precise in the paper), it is possible to generalize well even in some settings where the corresponding regression tasks would not generalize. Besides various technical challenges, it turns out that the key difference from the binary classification setting is that there are relatively fewer positive training examples of each class in the multiclass setting as the number of classes increases, making the multiclass problem “harder” than the binary one.

1 Introduction

Multiclass classification on standardized datasets is where the current deep-learning revolution really made the community take notice with previously unattainable levels of performance. Contemporary systems have demonstrated tremendous success at these tasks, typically using gigantic models with parameters that vastly exceed the (also large) number of data points used to train these models. In defiance of traditional statistical wisdom regarding overfitting, these big models can be trained to achieve zero training error even with noisy labels, but still generalize well in practice [84, 28].

To better understand this empirical phenomenon, one line of work uses appropriate high-dimensional linear models for regression problems to show how benign fitting of noise in training data is possible [31, 52, 4, 9, 55]. Essentially, the model must have enough "non-preferred" degrees of freedom to be able to absorb the training noise without contaminating predictions by too much. Simultaneously, there has to be enough of a preference for degrees of freedom that can capture the true pattern to enable it to survive the learning procedure and be well represented in the final learned model.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

A subsequent line of work studies binary classification [56, 13, 76] and shows that binary classification can generalize well beyond what can be proved by classical margin-based bounds [3] and there exist regimes where binary classification can even succeed in generalizing where regression fails — less preference is required for the degrees of freedom that capture the true pattern [56]. Very recently, the generalization of multiclass classification in similar models was studied in Wang et al. [77] but the analysis was limited to a fixed finite number of classes. In practice, we see that larger datasets often come with more classes and are tackled with even bigger models and so it is important to see what happens to generalization when everything scales together. To have a crisply understandable approach that allows everything to scale, this paper also adopts the bi-level covariance model with Gaussian features that is used in Muthukumar et al. [55, 56], Wang et al. [75], Wang et al. [77].

To understand classification, we must understand the role of training loss functions in determining what is learned. Empirical evidence shows that least-squares can yield classification performance competitive to cross-entropy minimization [64, 35, 11]. Muthukumar et al. [56], Hsu et al. [33] show that indeed with sufficient overparameterization, the support vector machine (SVM) solution, which also arises from minimizing the logistic loss using gradient descent [68, 36], is identical to that obtained by the minimum-norm interpolation (MNI) of binary labels — what would be obtained by gradient descent while minimizing the squared loss. A similar equivalence[1] holds for different variations of multiclass SVMs and the MNI of one-hot-encoded labels [77]. Consequently, this paper focuses on the MNI approach to overparameterized learning for multiclass classification.

2 Our contributions

Our study provides an asymptotic analysis of the error of the minimum-norm interpolating classifier for the multiclass classification problem with weighted Gaussian features. We consider an overparameterized setting using a bi-level feature weighting model where the number of features, classes, favored features, and the feature weights themselves all scale with the number of training points. Under this model, Theorem 5.1 provides sufficient conditions for good generalization in the form of a region in which as the number of training points increase, the number of classes grows slowly enough, the total number of features (i.e. level of overparameterization) grows fast enough, the number of favored features grows slowly enough, and the amount of favoring of those favored features is sufficient to allow for asymptotic generalization. We assume that our labels are generated noiselessly based on which of the first k features is the largest.[2]

To prove our main result, Theorem 5.1, we present a novel typicality-style argument featuring the feature margin (gap between the largest and second-largest feature) for computing sufficient conditions for correct classification utilizing the signal-processing inspired concepts of survival and contamination from Muthukumar et al. [55, 56] and leveraging the random-matrix analysis tools sharpened in Bartlett et al. [4]. The survival concept relates to the shrinkage induced by the regularizing effect of having lots of features in the context of min-norm interpolation — survival captures what is left of the true pattern after shrinkage. Contamination reflects the consequence of overparameterization when training via optimization: in addition to the true pattern, there is an infinite family of other[3] false patterns (aliases) that also happen to explain the limited training data, and the optimizer ends up hedging its bet across the true pattern and these other competing false explanations. The learned false patterns contaminate the predictions on test points, and this can be quantified by the relevant standard deviation.

The key is analyzing what happens with multiclass training data where there are relatively fewer positive examples of each class, and where the training data for a particular class is not independent of the features corresponding to other classes. The analysis shows that as a result of having fewer positive exemplars for a class relative to the total size of the training data, the survival drops by a factor of k (the number of classes), while the contamination only drops by a factor of $\sqrt{k}$. As in binary classification, the ratio of the relevant survival to contamination terms plays the role of the effective signal-to-noise ratio and shows up as a key quantity in our error analysis (Equation (22) from

[1] For an interesting alternative perspective on this equivalence as an indication of a potential bug instead of as a promising feature, see Shamir [67].

[2] This assumption is without loss of generality for the bi-level model as long as the classes are defined by orthogonal directions as in Wang et al. [77].

[3] This is related to what is called the challenge of "underspecification" in ML [19], and this in turn is also one aspect of the challenge of covariate shifts [73].

Section 5.1). When this ratio grows asymptotically to ∞, multiclass classification generalizes well. To the best of our knowledge, this is the first work that quantifies this effect of fewer informative samples per class and in what sense that makes multiclass classification harder than binary classification. The closest related work ([77]) only considers multiclass classification in the fixed finite class setting and consequently, doesn't compute exact dependencies on the number of classes k. We provide a more detailed comparison of our work with Wang et al. [77] and Muthukumar et al. [56] in Appendix H of the Supplemental material.

3 Related Work

The present work is situated within a larger stream of theoretical research trying to understand why overparameterized learning works and its limits. The limited page budget here forces brevity, but we recommend the recent surveys Bartlett et al. [5], Belkin [6], Dar et al. [20] for further context.

Classically, by either operating in the underparameterized regime or by performing explicit regularization, we can force the training procedure to average out the harmful effects of training noise and thereby hope to obtain good generalization. The present cycle of seeking a deeper understanding began after it was observed that modern deep networks were overparameterized, capable of memorizing noise, and yet still generalized well, even when they were trained without explicit regularization [59, 84]. Experiments in Geiger et al. [28], Belkin et al. [8] observed a double-descent behavior of the generalization error where in addition to the traditional U-shaped curve in the underparameterized regime, the error decreases in the overparameterized regime as we increase the number of model parameters. This double descent phenomenon is not unique to deep learning models and was replicated for kernel learning [7]. Further, the good generalization performance in the overparameterized regime cannot be explained by traditional worst-case generalization bounds based on Rademacher complexity or VC-dimension since the models have the capacity to fit purely random labels. Overparameterized models must therefore have some fortuitous combination of the model architecture with the training algorithm that leads us to a particular solution that generalizes well.

To understand the phenomenon better, several works study the simpler setting of overparameterized linear regression.The minimum-ℓ_2 norm[4] interpolator is of particular interest since gradient descent on the squared loss has an implicit[5] bias towards this solution in the overparameterized regime [24] and has been studied extensively. (An incomplete list is Hastie et al. [31], Mei and Montanari [52], Bartlett et al. [4], Belkin et al. [9], Muthukumar et al. [55], Bibas et al. [10], Kobak et al. [41], Wu and Xu [81], Richards et al. [63].) To generalize well, the underlying feature family must satisfy a balance between having a few important directions that sufficiently favor the true pattern, and a large number of unimportant directions that can absorb the noise in a harmless manner.

3.1 High dimensional binary classification

Both concurrently with and subsequent to the wave of analyses on overparameterized regression, researchers turned their attention to binary classification. A line of work poses the overparameterized binary classification problem as an optimization problem and analyzes it directly to obtain precise asymptotic behaviours of the generalization error [22, 66, 37, 69, 54, 38, 70]. The key technical tool employed in these works is the Convex Gaussian Min-max Theorem and the resultant error formulas involve solutions to a system of non-linear equations that typically do not admit closed-form expressions. The generalization error of the max-margin SVM has also been analyzed directly by studying the iterates of gradient descent in [13] and leveraging the implicit regularization perspective of optimization algorithms.

However, although the above works did significantly enhance our understanding of binary classification in the overparameterized regime, a fundamental question was not answered: "Is classification easier than regression?" While the classification task is easier than the regression task at test time (regression requires us to correctly predict a real value while binary classification requires us to only predict its sign correctly), the training data for classification is less informative than that for regression

[4] The minimum-ℓ_1 norm interpolator has also been studied in Muthukumar et al. [55], Mitra [53], Li and Wei [50], Wang et al. [75] and while sparsity-seeking behavior helps preserve the true signal (if the true pattern indeed depends only on a few features), it poses a challenge for the harmless absorption of noise since the desired averaging behaviour is not achieved fully [55].

[5] In fact, there is an important complementary literature that brings out the implicit regularization performed by training methods, especially variants of gradient descent and stochastic gradient descent, and how the underlying architecture of the model shapes this implicit regularization [30, 68, 36, 80, 57, 2, 82].

since the labels are also binary. As described earlier, this question was answered in Muthukumar et al. [56], by exhibiting an asymptotic regime where binary classification error goes to zero, but the regression error does not. This was shown using Gaussian features with a bi-level covariance model. It turns out that the level of anisotropy (favoring of true features) required to perform regression correctly is significantly higher than that required for binary classification.

The key to the result in Muthukumar et al. [56] was the signal-processing inspired survival/contamination framework introduced in Muthukumar et al. [55] as a reconceptualization of the "effective ranks" perspective of Bartlett et al. [4]. For binary classification to succeed, what matters is that the survival exceed the contamination so that the sign of the prediction remains correct. Meanwhile, regression is harder since for regression to succeed, the survival must also tend to 1.

3.2 Multiclass classification and the role of training loss function

There is a large classical body of work on multiclass classification algorithms [79, 12, 23, 18, 46], with further works giving computationally efficient algorithms for extreme multiclass problems with a huge number of classes [15, 83, 62]. Numerous theoretical works investigate the consistency of classifiers [85, 60, 61, 71, 14]. Finite-sample analysis of the generalization error in multiclass classification problems in the underparameterized regime has been studied in Koltchinskii and Panchenko [42], Guermeur [29], Allwein et al. [1], Li et al. [49], Cortes et al. [16], Lei et al. [47], Maurer [51], Lei et al. [48], Kuznetsov et al. [44, 45] and includes both data dependent bounds using Rademacher complexity, Gaussian complexity and covering numbers as well as data-independent bounds using the VC dimension. Recent work [72] leverages the Convex Gaussian Min-max Theorem to precisely characterize the asymptotic behaviour of the least-squares classifier in underparameterized multiclass classification.

So, how different is multiclass classification from binary classification? The test time task is more difficult and for the same total number of training points, we have fewer positive training examples from each class. Several empirical studies comparing the performances of multiclass classification via learning multiple binary classifiers have been undertaken [64, 25, 1]. The effects of the loss function while using deep nets to perform classification has also been investigated [32, 26, 43, 11, 21, 40, 35, 39]. Empirical evidence of least-squares minimization yielding competitive test classification performance to cross-entropy minimization has been presented in Rifkin and Klautau [64], Hui and Belkin [35], Bosman et al. [11].

More recently, Wang et al. [77] makes progress towards bridging the gap between empirical observations and theoretical understanding by proving that in certain overparameterized regimes the solution to a multiclass SVM problem is identical to the one obtained by minimum-norm interpolation of one-hot encoded labels (equivalently, that gradient descent on squared loss leads to the same solution as gradient descent on cross-entropy loss as a result of implicit bias of these algorithms [24, 36, 68]). In addition, Wang et al. [77] extends the analysis presented in Muthukumar et al. [56] for the binary classification problem to the multiclass problem with finitely many classes via an interesting reduction to analyzing a finite set of pairwise competitions, all of which must be won for multiclass classification to succeed. (We give further comments on the relationship of the present paper with Wang et al. [77] in Appendix H of the Supplemental material.)

4 Problem setup

We consider the multiclass classification problem with k classes. The training data consists of n pairs $\{\mathbf{x}_i, \ell_i\}_{i=1}^n$ where $x_i \in \mathbb{R}^d$ are i.i.d Gaussian vectors drawn from distribution,

$$\mathbf{x}_i \sim \mathcal{N}(0, I_d). \tag{1}$$

We make the following assumption on how the labels $\ell_i \in [k]$ are generated.

Assumption 4.1. ***Orthogonal classes noiseless model***[6] *The class labels ℓ_i are generated based on which of the first k dimensions of a point $\mathbf{x}_i$ has the largest value,*

$$\ell_i = \operatorname*{argmax}_{m \in [k]} \mathbf{x}_i[m]. \tag{2}$$

[6] A more generic model is $\ell_i = \operatorname{argmax}_{m\in[k]} \boldsymbol{\mu}_m^\top \mathbf{x}_i$ where the $\boldsymbol{\mu}_m$ are unit norm orthogonal vectors. If we further assume the bi-level model(Definition 4.2) and that the vectors $\boldsymbol{\mu}_m$ have no support outside of the favored features then it suffices to consider the simplified setting where $\boldsymbol{\mu}_m$ are 1-sparse unit vectors like we do here, due to the indifference of minimum norm interpolation to orthogonal transformations.

We use the notation $x_i[m]$ to refer to the m^{th} element of vector $\mathbf{x}_i$. For clarity of exposition, we make explicit a feature weighting that transforms the training points as follows:

$$x_i^w[j] = \sqrt{\lambda_j} x_i[j] \quad \forall j \in [d]. \tag{3}$$

Here $\boldsymbol{\lambda} \in \mathbb{R}^d$ contains the squared feature weights. The feature weighting serves the role of favoring the true pattern, something that is essential for good generalization.[7]

The weighted feature matrix $\mathbf{X}^w \in \mathbb{R}^{n \times d}$ is given by,

$$\mathbf{X}^w = \begin{bmatrix} \mathbf{x}_1^w & \dots & \mathbf{x}_j^w & \dots & \mathbf{x}_n^w \end{bmatrix}^\top = \begin{bmatrix} \sqrt{\lambda_1}\mathbf{z}_1 & \dots & \sqrt{\lambda_j}\mathbf{z}_j & \dots & \sqrt{\lambda_d}\mathbf{z}_d \end{bmatrix}, \tag{4}$$

where $\mathbf{z}_j \in \mathbb{R}^n$ contains the j^{th} features from the n training points. Note that $\mathbf{z}_j \sim \mathcal{N}(0, I_n)$ are i.i.d Gaussians. We use a one-hot encoding for representing the labels as the matrix $\mathbf{Y}^{oh} \in \mathbb{R}^{n \times k}$,

$$\mathbf{Y}^{oh} = \begin{bmatrix} \mathbf{y}_1^{oh} & \dots & \mathbf{y}_m^{oh} & \dots & \mathbf{y}_k^{oh} \end{bmatrix}, \tag{5}$$

where,

$$y_m^{oh}[i] = \begin{cases} 1, & \text{if } \ell_i = m \\ 0, & \text{otherwise} \end{cases}. \tag{6}$$

A zero-mean variant of the encoding where we subtract the mean $\frac{1}{k}$ from each entry is denoted:

$$\mathbf{y}_m = \mathbf{y}_m^{oh} - \frac{1}{k}\mathbf{1}. \tag{7}$$

Our classifier consists of k coefficient vectors $\hat{\mathbf{f}}_m$ for $m \in [k]$ that are learned by minimum-norm interpolation of the zero-mean one-hot variants using the weighted features.[8]

$$\hat{\mathbf{f}}_m = \arg\min_{\mathbf{f}} \|\mathbf{f}\|_2 \tag{8}$$

$$\text{s.t. } \mathbf{X}^w \mathbf{f} = \mathbf{y}_m^{oh} - \frac{1}{k}\mathbf{1}. \tag{9}$$

We can express these coefficients in closed form as,

$$\hat{\mathbf{f}}_m = (\mathbf{X}^w)^\top \left(\mathbf{X}^w (\mathbf{X}^w)^\top\right)^{-1} \mathbf{y}_m. \tag{10}$$

On a test point $\mathbf{x}_{test} \sim \mathcal{N}(0, I_d)$ we predict a label as follows: First, we transform the test point into the weighted feature space to obtain $\mathbf{x}_{test}^w$ where $x_{test}^w[j] = \sqrt{\lambda_j} x_{test}[j]$ for $j \in [d]$. Then we compute k scalar "scores" and assign the class based on the largest score as follows:

$$\hat{\ell} = \underset{1 \le m \le k}{\operatorname{argmax}} \, \hat{\mathbf{f}}_m^\top \mathbf{x}_{test}^w. \tag{11}$$

The true label of the test point is $\ell_{test} = \operatorname{argmax}_{1 \le m \le k} x_{test}[m]$. A misclassification event $\mathcal{E}_{err}$ occurs iff

$$\underset{1 \le m \le k}{\operatorname{argmax}} \, x_{test}[m] \neq \underset{1 \le m \le k}{\operatorname{argmax}} \, \hat{\mathbf{f}}_m^\top \mathbf{x}_{test}^w. \tag{12}$$

In our work we determine sufficient conditions under which the probability of misclassification (computed over the randomness in both the training data and test point) goes to zero in an asymptotic regime where the number of training points, number of features, number of classes and feature weights scale according to the bi-level ensemble model.

[7] Our weighted feature model is equivalent to the one used in other works (e.g. [56]) that assume that the covariates come from an anisotropic Gaussian with a covariance matrix that favors the truly important directions.

[8] The classifier learned via this method is equivalent to those obtained by other natural training methods under sufficient overparameterization [77].

Definition 4.2. *(**Bi-level ensemble**): The bi-level ensemble is parameterized by p, q, r and t where $p > 1$, $0 \leq r < 1$, $0 < q < (p - r)$ and $0 \leq t < r$. Here, parameter p controls the extent of overparameterization, r determines the number of favored features, q controls the weights on favored features and t controls the number of classes. The number of features (d), number of favored features (s), number of classes (k) and feature weights $(\sqrt{\lambda_j})$ all scale with the number of training points (n) as follows:*

$$d = \lfloor n^p \rfloor, s = \lfloor n^r \rfloor, a = n^{-q}, k = c_k \lfloor n^t \rfloor, \tag{13}$$

where c_k is a positive integer. The feature weights are given by,

$$\sqrt{\lambda_j} = \begin{cases} \sqrt{\frac{ad}{s}}, & 1 \leq j \leq s \\ \sqrt{\frac{(1-a)d}{d-s}}, & \text{otherwise} \end{cases}. \tag{14}$$

We provide a visualization of the bi-level model in Figure 1.

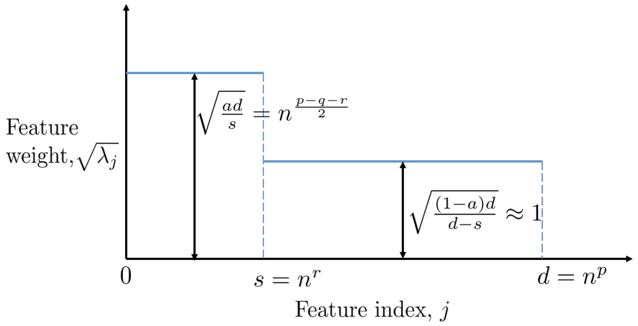

Feature weight, $\sqrt{\lambda_j}$
$\sqrt{\frac{ad}{s}} = n^{\frac{p-q-r}{2}}$
$\sqrt{\frac{(1-a)d}{d-s}} \approx 1$
0
$s = n^r$
$d = n^p$
Feature index, j

Figure 1: Bi-level feature weighting model. The first s features have a higher weight and are favored during minimum-norm interpolation. These can be thought of as the square-roots of the eigenvalues of the feature covariance matrix in a Gaussian model for the covariates as in Bartlett et al. [4].

5 Main result

Theorem 5.1. *(**Asymptotic classification region in the bi-level model**): Under the bi-level ensemble model 4.2, when the true data generating process is 1-sparse (Assumption 4.1), the probability of misclassification $P(\mathcal{E}_{err}) \to 0$ as $n \to \infty$ if the following conditions hold:*

$$t < \min\left(r, 1 - r, p + 1 - 2(q + r), p - 2, 2q + r - 2\right) \tag{15}$$

$$q + r > 1. \tag{16}$$

Note that from Muthukumar et al. [56], the condition $q + r > 1$ corresponds to the regime where the corresponding regression does not generalize well and thus our result shows that multiclass classification can generalize in regimes where the corresponding regression problem does not. In this challenging regime, the empirical eigenstructure does not reveal the true nature of underlying features as illustrated in Appendix J.

Figure 2 visualizes the regimes by considering slices of the four dimensional scaling parameter space of p, q, r and t. (1a) and (2a) fix the value of q to 0.75 and 0.95 respectively and contrast the multiclass problem with a fixed finite number of classes ($t = 0$) to the binary classification and regression problems. From these plots we observe that if we fix p, q, t and increase r, i.e. increasing how many features are favored (and thereby favoring each of them less), we transition from the regime where both regression and binary classification work, into the regime where binary classification works but regression does not, then the regime where this paper can prove multiclass classification works and finally to the regime where neither regression nor binary classification works.

In Figure 2, subplots (1b),(1c),(2b) and (2c) each visualize a slice along the r and t (class scaling) dimensions with fixed p and q. The x axis itself in these plots corresponds to a fixed finite classes setting. From (1b) we observe that the right-hand boundary of the region where multiclass classification generalizes well contains two slopes. These slopes arise from the two conditions $t < 1 - r$ and

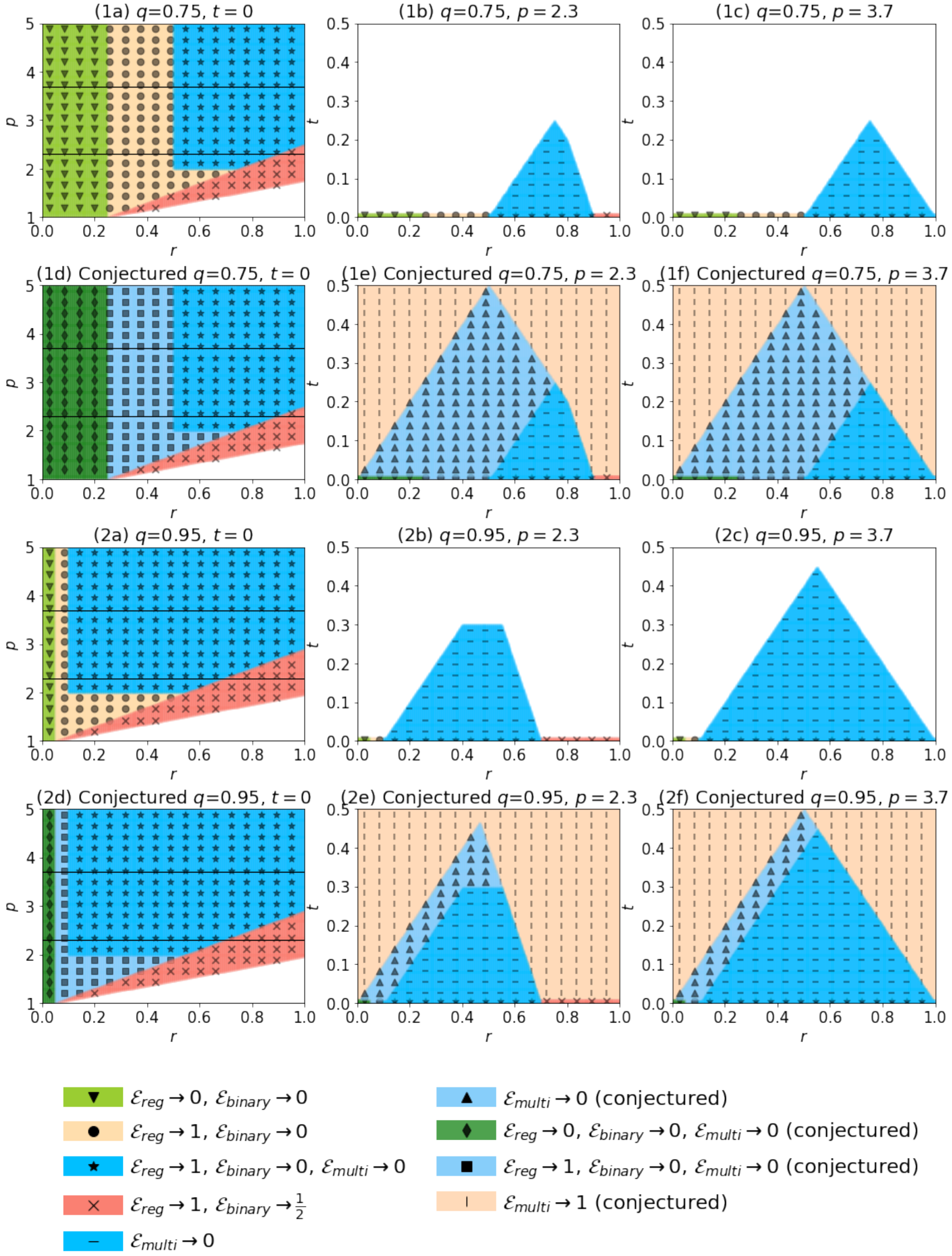


Figure 2: Visualization of the bi-level regimes in four dimensions p, q, r, t. (1a) and (2a) contrast multiclass classification with finite classes to binary classification and regression. The horizontal lines $p = 2.3$ and $p = 3.7$ correspond to the slices visualized in (1b), (1c), (2b) and (2c). The conjectured regimes are visualized in (1d), (1e), (1f), (2d), (2e) and (2f).

$t < p + 1 - 2(q + r)$ in Theorem 5.1 and are a result of either contamination from favored (but not true) features dominating or contamination from the unfavored features dominating. In (1c) we are in the regime where binary classification works for all values of $r < 1$. However, as we increase t, eventually multiclass classification stops working.[9]

When we go from the binary problem to a multiclass problem with k classes, the survival drops by a factor of k as a consequence of having only $\frac{1}{k}$ fraction of positive training examples per class. This is because the one-hot labels we interpolate while training have fewer large values close to 1 that are able to positively correlate with the true feature vector. Having fewer positive exemplars also reduces the total energy in the training vector by a factor of k, and because of the square-root relationship of the standard deviation to the energy, the contamination only shrinks by a factor of $\sqrt{k}$. The overall survival/contamination ratio decreases by a factor of $\sqrt{k}$ making the multiclass classification task more difficult.[10] An interesting observation here is the amount of favoring required for good generalization is linked to the number of positive training examples per class. Indeed, if we consider a setting where the binary classification problem generalizes well, and we switch to the k class multiclass problem, then by increasing the number of training samples k fold (and thus matching the number of positive training examples per class in the multiclass case to the binary case) and keeping the number of features and feature weights constant we can generalize well for multiclass classification. (Appendix G of the Supplemental material elaborates on this phenomenon, as well as why it is somewhat surprising.)

Next, we present a brief overview of our proof that utilizes the survival/contamination analysis framework from Muthukumar et al. [56] along with a typicality-inspired argument where the feature margin (difference between largest and second largest feature) on the test point plays a key role. The complete proof is provided in Appendices B, C, D, and E of the Supplemental material.

5.1 Proof sketch

Assume without loss of generality that for the test point $\mathbf{x}_{test} \sim \mathcal{N}(0, I_d)$, the true class is α for some $\alpha \in [k]$. Let $\mathbf{x}^w_{test}$ be the weighted version of this test point. A necessary and sufficient condition for classification error is that for some $\beta \neq \alpha, \beta \in [k]$,

$$\begin{aligned}\widehat{f}_\alpha[\alpha]x^w_{test}[\alpha] + \widehat{f}_\alpha[\beta]x^w_{test}[\beta] + \sum_{j\notin\{\alpha,\beta\}} \widehat{f}_\alpha[j]x^w_{test}[j] < \widehat{f}_\beta[\alpha]x^w_{test}[\alpha] \\ + \widehat{f}_\beta[\beta]x^w_{test}[\beta] + \sum_{j\notin\{\alpha,\beta\}} \widehat{f}_\beta[j]x^w_{test}[j].\end{aligned} \tag{17}$$

By converting into the unweighted feature space we obtain

$$\lambda_\alpha \widehat{h}_{\alpha,\beta}[\alpha]x_{test}[\alpha] - \lambda_\beta \widehat{h}_{\beta,\alpha}[\beta]x_{test}[\beta] < \sum_{j\notin\{\alpha,\beta\}} \lambda_j \widehat{h}_{\beta,\alpha}[j]x_{test}[j], \tag{18}$$

where

$$\widehat{h}_{\alpha,\beta}[j] = \lambda_j^{-1/2}(\hat{f}_\alpha[j] - \hat{f}_\beta[j]). \tag{19}$$

Performing some algebraic manipulations and because $\lambda_\alpha = \lambda_\beta = \lambda$ since both α and β are favored features, we can rewrite this as

$$\begin{aligned}\frac{\lambda\widehat{h}_{\alpha,\beta}[\alpha]}{\mathsf{CN}_{\alpha,\beta}}\left((x_{test}[\alpha] - x_{test}[\beta]) + x_{test}[\beta]\frac{\widehat{h}_{\alpha,\beta}[\alpha] - \widehat{h}_{\beta,\alpha}[\beta]}{\widehat{h}_{\alpha,\beta}[\alpha]}\right) \\ < \frac{1}{\mathsf{CN}_{\alpha,\beta}} \sum_{j\notin\{\alpha,\beta\}} \lambda_j \widehat{h}_{\beta,\alpha}[j]x_{test}[j],\end{aligned} \tag{20}$$

[9] To be precise, what the region actually illustrates is that our proof approach stops being able to show that multiclass classification works. In the Conclusion section, we conjecture where we believe that multiclass classification actually stops working. The conjectured regions are illustrated in (1e),(1f),(2e) and (2f).

[10] This is also responsible for contamination due to favored features being able to cause errors. For binary classification, because the true feature survival is constant (depending only on the level of label noise), the survival can always asymptotically overcome any contamination from other favored features [56].

where

$$\mathsf{CN}_{\alpha,\beta} = \sqrt{\left(\sum_{j\notin\{\alpha,\beta\}} \lambda_j^2 (\widehat{h}_{\beta,\alpha}[j])^2\right)}. \tag{21}$$

We divide by $\mathsf{CN}_{\alpha,\beta}$ to normalize the RHS above to have a standard normal distribution. Next, by removing the dependency on β, we obtain a sufficient condition for correct classification:

$$\underbrace{\frac{\min_\beta \lambda \widehat{h}_{\alpha,\beta}[\alpha]}{\max_\beta \mathsf{CN}_{\alpha,\beta}}}_{\text{SU/CN ratio}} \left(\underbrace{\min_\beta \left(x_{test}[\alpha] - x_{test}[\beta]\right)}_{\text{closest feature margin}} - \underbrace{\max_\beta |x_{test}[\beta]|}_{\text{largest competing feature}} \cdot \underbrace{\max_\beta \left| \frac{\widehat{h}_{\alpha,\beta}[\alpha] - \widehat{h}_{\beta,\alpha}[\beta]}{\widehat{h}_{\alpha,\beta}[\alpha]} \right|}_{\text{survival variation}} \right)$$
$$> \underbrace{\max_\beta \frac{1}{\mathsf{CN}_{\alpha,\beta}} \left(\sum_{j\notin\{\alpha,\beta\}} \lambda_j \widehat{h}_{\beta,\alpha}[j] x_{test}[j] \right)}_{\text{normalized contamination}}. \tag{22}$$

Here the min and max are over all competing features: $1 \leq \beta \leq k, \beta \neq \alpha$ and the sum is over all d feature indices except α and β, but we simplify the notation for convenience. We show via intermediate lemmas introduced in Appendix B of the Supplemental material that under the conditions specified in Theorem 5.1, with sufficiently high probability[11], the relevant survival to contamination SU/CN ratio grows at a polynomial rate n^v for some $v > 0$, the closest feature margin shrinks at a less-than-polynomial rate $1/\sqrt{\ln nk}$, the survival variation decays at a polynomial rate n^{-u} for some $u > 0$. Further, the magnitudes of the largest competing feature and the normalized contamination are no more than $\sqrt{\ln(nk)}$.

This implies that the left-hand side of Equation (22) grows at a polynomial rate n^v (ignoring logarithmic terms) and dominates the right-hand side which grows at the much slower rate $\sqrt{\ln nk}$. A survival/contamination ratio also plays a key role in the analysis of the binary classification problem in Muthukumar et al. [56] but in the multiclass setting, we additionally have the survival variation term and feature margin playing important roles since we are comparing different scores while predicting the class label. For correct classification, the survival/contamination ratio must be sufficiently large, the survival variation must be small enough and the feature margin must be sufficiently large.

6 Conclusion

In this work we compute sufficient conditions for good generalization of multiclass classification in a bi-level overparameterized linear model with Gaussian features. We observed that multiclass classification can generalize even when the regression problem does not generalize (for $q + r > 1$). Further, the multiclass problem is "harder" than the binary problem because we have fewer positive training examples per class. The nature of the training data complicates our analysis in the multiclass setting since the true class labels are generated by comparing k features and thus we no longer have independence of the encoded class label y with any of these features. This becomes relevant when we compute bounds on the survival and contamination quantities since the Hanson-Wright inequality [65] is no longer applicable directly on the quantities of interest as was the case for the binary classification problem in prior work [56]. As a consequence of working around this non-independence we believe that our sufficient conditions for good generalization in the regime $q + r > 1$ are loose.

Even though in our work we focus on the regime where regression does not work, $q + r > 1$, we can extend the analysis to the regime where $q + r < 1$ by grinding through the expressions for survival and contamination in this regime. Even in this regime, for multiclass training data, survival is of the order $\frac{1}{k}$ while contamination scales similarly to the regime $q + r > 1$. Thus, while it is true that for

[11] This is where we leverage the idea of typicality-style proofs in information theory [17] to avoid unnecessarily loose union bounds that end up being dominated by the atypical behavior of quantities. In our case, by pulling the feature margin out explicitly, we can just deal with its typical behavior. Similarly, the typical behavior of the largest competing feature and the true feature is all that matters.

binary classification or a fixed number of classes, the regime where regression works is a regime where classification also works, this need not be true if there are too many classes.

We conjecture that the following is a set of necessary and sufficient conditions for asymptotically good generalization (We elaborate on this in Appendix F in the Supplemental material):

Conjecture 6.1. ***(Conjectured bi-level regions)****: Under the bi-level ensemble model 4.2, when the true data generating process is 1-sparse (Assumption 4.1), as $n \to \infty$, the probability of misclassification event $P(\mathcal{E}_{err})$ behaves as follows:*

$$P(\mathcal{E}_{err}) \to \begin{cases} 0, & \text{if } t < \min\left(r, 1-r, p+1-2\cdot\max(1, q+r)\right) \\ 1, & \text{if } t > \min\left(r, 1-r, p+1-2\cdot\max(1, q+r)\right) \end{cases}. \tag{23}$$

The conjectured regions are visualized in (1d),(1e),(1f),(2d),(2e) and (2f) in Figure 2. Subfigures (1d) and (2d) illustrate that we believe multiclass classification with finitely many classes works if binary classification works. Further, comparing (1e) to (2e) when we increase q, the conjectured parameter region where multiclass classification works shrinks since we decrease the amount of favoring of true features. Interestingly, the nature of the looseness in our approach is such that our proof technique is able to recover a larger fraction of the conjectured region for larger q which intuitively is a result of less favoring leading to stronger concentration of certain random quantities. Tightening the potential looseness in our analysis and proving the converse result by computing sufficient conditions for poor generalization of multiclass classification are interesting avenues of future work.

Further, although the present analysis focuses on solutions that exactly interpolate the training data, we can extend our results to account for additional ridge regularization by viewing ridge regularization as minimum-norm interpolation using augmented contamination-free features as in the Appendix of Muthukumar et al. [55] and computing bounds leveraging tools from Tsigler and Bartlett [74]. Our assumption of the strict bi-level weighting model is largely to simplify the calculations and by substituting terms appropriately in our lemmas from Appendix B in the Supplemental material, it should be possible to compute results for other weighting models. Finally, exploring the new phenomena that can be encountered as we go beyond the 1-sparse noiseless model is an exciting direction for future work.

Acknowledgments and Disclosure of Funding

We are grateful to our earlier collaborators Vidya Muthukumar, Misha Belkin, Daniel Hsu, and Adhyyan Narang. In addition, we want to thank the students and course staff for the Fall 2020 iteration of Berkeley's CS189/289A machine learning courses, where we had adapted ideas from Muthukumar et al. [55, 56] in teaching the foundations of modern machine learning — the need for the present paper became more clear during this process.

We gratefully acknowledge the support from ML4Wireless center member companies and NSF grants AST-2132700 and AST-2037852 for making this research possible.

References

[1] Erin Allwein, Robert E. Schapire, and Yoram Singer. Reducing multiclass to binary: A unifying approach for margin classifiers. Journal of Machine Learning Research, 1:113–141, 2000.

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

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

[4] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. Proceedings of the National Academy of Sciences, 117(48):30063–30070, 2020.

[5] Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. Acta numerica, 30:87–201, 2021.

[6] Mikhail Belkin. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. Acta Numerica, 30:203–248, 2021.

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

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

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

[10] Koby Bibas, Yaniv Fogel, and Meir Feder. A new look at an old problem: A universal learning approach to linear regression. CoRR, abs/1905.04708, 2019. URL `http://arxiv.org/abs/1905.04708`.

[11] Anna Bosman, Andries Engelbrecht, and Mardé Helbig. Visualising basins of attraction for the cross-entropy and the squared error neural network loss functions. Neurocomputing, 400, 03 2020. doi: 10.1016/j.neucom.2020.02.113.

[12] Erin J. Bredensteiner and Kristin P. Bennett. Multicategory classification by support vector machines. Computational Optimization and Applications, 12, 1999. doi: 10.1023/A:1008663629662.

[13] Niladri S Chatterji and Philip M Long. Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. Journal of Machine Learning Research, 22(129):1–30, 2021.

[14] Di-Rong Chen and Tao Sun. Consistency of multiclass empirical risk minimization methods based on convex loss. Journal of Machine Learning Research, 7:2435–2447, dec 2006. ISSN 1532-4435.

[15] Anna Choromanska, Alekh Agarwal, and John Langford. Extreme multi class classification. In NIPS Workshop: eXtreme Classification, submitted, volume 1, pages 2–1, 2013.

[16] Corinna Cortes, Vitaly Kuznetsov, Mehryar Mohri, and Scott Yang. Structured prediction theory based on factor graph complexity. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 29. Curran Associates, Inc., 2016. URL `https://proceedings.neurips.cc/paper/2016/file/535ab76633d94208236a2e829ea6d888-Paper.pdf`.

[17] Thomas M. Cover and Joy A. Thomas. Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing). Wiley-Interscience, July 2006. ISBN 0471241954.

[18] Koby Crammer, Yoram Singer, Nello Cristianini, John Shawe-taylor, and Bob Williamson. On the algorithmic implementation of multiclass kernel-based vector machines. Journal of Machine Learning Research, 2:265–292, 2001.

[19] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. arXiv preprint arXiv:2011.03395, 2020.

[20] Yehuda Dar, Vidya Muthukumar, and Richard G Baraniuk. A farewell to the bias-variance tradeoff? an overview of the theory of overparameterized machine learning. arXiv preprint arXiv:2109.02355, 2021.

[21] Ahmet Demirkaya, Jiasi Chen, and Samet Oymak. Exploring the role of loss functions in multiclass classification. In 2020 54th Annual Conference on Information Sciences and Systems (CISS), pages 1–5, 2020. doi: 10.1109/CISS48834.2020.1570627167.

[22] Zeyu Deng, Abla Kammoun, and Christos Thrampoulidis. A model of double descent for high-dimensional binary linear classification. Information and Inference: A Journal of the IMA, 04 2021. ISSN 2049-8772. doi: 10.1093/imaiai/iaab002. URL https://doi.org/10.1093/imaiai/iaab002. iaab002.

[23] Thomas G Dietterich and Ghulum Bakiri. Solving multiclass learning problems via error-correcting output codes. Journal of Artificial Intelligence Research, 2(1):263–286, 1994. ISSN 1076-9757.

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

[25] Johannes Fürnkranz. Round robin classification. Journal of Machine Learning Research, 2: 721–747, 2002.

[26] Krzysztof Gajowniczek, Leszek Chmielewski, Arkadiusz Orłowski, and Tomasz Ząbkowski. Generalized entropy cost function in neural networks. In International Conference on Artificial Neural Networks, pages 128–136, 10 2017. ISBN 978-3-319-68611-0. doi: 10.1007/978-3-319-68612-7_15.

[27] Robert G Gallager. Information theory and reliable communication, volume 588. Springer, 1968.

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

[29] Yann Guermeur. Combining Discriminant Models with New Multi-Class SVMs. Pattern Anal. Appl., 5:168–179, 06 2002. doi: 10.1007/s100440200015.

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

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

[32] Le Hou, Chen-Ping Yu, and Dimitris Samaras. Squared Earth Mover's Distance-based Loss for Training Deep Neural Networks. arXiv e-prints, art. arXiv:1611.05916, November 2016.

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

[34] Iosif Pinelis (https://mathoverflow.net/users/36721/iosif pinelis). Concentration and anti-concentration of gap between largest and second largest value in gaussian iid sample. MathOverflow. URL https://mathoverflow.net/q/379688. URL:https://mathoverflow.net/q/379688 (version: 2020-12-25).

[35] Like Hui and Mikhail Belkin. Evaluation of Neural Architectures Trained with Square Loss vs Cross-Entropy in Classification Tasks. arXiv e-prints, art. arXiv:2006.07322, June 2020.

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

[37] Abla Kammoun and Mohamed-Slim AlouiniFellow. On the precise error analysis of support vector machines. IEEE Open Journal of Signal Processing, 2:99–118, 2021.

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

[39] Ganesh Ramachandra Kini, Orestis Paraskevas, Samet Oymak, and Christos Thrampoulidis. Label-imbalanced and group-sensitive classification under overparameterization. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, Advances in Neural Information Processing Systems, volume 34, pages 18970–18983. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/9dfcf16f0adbc5e2a55ef02db36bac7f-Paper.pdf.

[40] Doug M. Kline and Victor L. Berardi. Revisiting squared-error and cross-entropy functions for training neural network classifiers. Neural Computing & Applications, 14:310–318, 2005.

[41] Dmitry Kobak, Jonathan Lomond, and Benoit Sanchez. The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization. Journal of Machine Learning Research, 21:169–1, 2020.

[42] Vladimir Koltchinskii and Dmitry Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. The Annals of Statistics, 30(1):1–50, 2002. ISSN 00905364. URL http://www.jstor.org/stable/2700001.

[43] Himanshu Kumar and P. Shanti Sastry. Robust loss functions for learning multi-class classifiers. 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pages 687–692, 2018.

[44] Vitaly Kuznetsov, Mehryar Mohri, and Umar Syed. Multi-class deep boosting. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/7bb060764a818184ebb1cc0d43d382aa-Paper.pdf.

[45] Vitaly Kuznetsov, Mehryar Mohri, and Umar Syed. Rademacher complexity margin bounds for learning with a large number of classes. In ICML Workshop on Extreme Classification: Learning with a Very Large Number of Labels, 2015.

[46] Yoonkyung Lee, Yi Lin, and Grace Wahba. Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. Journal of the American Statistical Association, 99(465):67–81, 2004. doi: 10.1198/016214504000000098. URL https://doi.org/10.1198/016214504000000098.

[47] Yunwen Lei, Urun Dogan, Alexander Binder, and Marius Kloft. Multi-class SVMs: From Tighter Data-Dependent Generalization Bounds to Novel Algorithms. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/3a029f04d76d32e79367c4b3255dda4d-Paper.pdf.

[48] Yunwen Lei, Ürün Dogan, Ding-Xuan Zhou, and Marius Kloft. Data-dependent generalization bounds for multi-class classification. IEEE Transactions on Information Theory, 65(5):2995–3021, 2019. doi: 10.1109/TIT.2019.2893916.

[49] Jian Li, Yong Liu, Rong Yin, Hua Zhang, Lizhong Ding, and Weiping Wang. Multi-class learning: From theory to algorithm. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, Advances in Neural Information Processing Systems, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/1141938ba2c2b13f5505d7c424ebae5f-Paper.pdf.

[50] Yue Li and Yuting Wei. Minimum ℓ_1-norm interpolators: Precise asymptotics and multiple descent. arXiv preprint arXiv:2110.09502, 2021.

[51] Andreas Maurer. A vector-contraction inequality for Rademacher complexities. In Hans Ulrich Simon Ronald Ortner and Sandra Zilles, editors, Algorithmic Learning Theory, pages 3–17. Springer International Publishing, 2016.

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

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

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

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

[56] Vidya Muthukumar, Adhyyan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel J. Hsu, and Anant Sahai. Classification vs regression in overparameterized regimes: Does the loss function matter? Journal of Machine Learning Research, 22:222:1–222:69, 2021.

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

[58] Preetum Nakkiran. More Data Can Hurt for Linear Regression: Sample-wise Double Descent. arXiv e-prints, art. arXiv:1912.07242, December 2019.

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

[60] Bernardo Ávila Pires and Csaba Szepesvári. Multiclass Classification Calibration Functions. arXiv e-prints, art. arXiv:1609.06385, September 2016.

[61] Bernardo Ávila Pires, Mohammad Ghavamzadeh, and Csaba Szepesvári. Cost-sensitive multiclass classification risk bounds. In Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28, pages 1391–1399, 2013.

[62] Ankit Singh Rawat, Jiecao Chen, Felix Xinnan X Yu, Ananda Theertha Suresh, and Sanjiv Kumar. Sampled softmax with random fourier features. Advances in Neural Information Processing Systems, 32, 2019.

[63] Dominic Richards, Jaouad Mourtada, and Lorenzo Rosasco. Asymptotics of ridge (less) regression under general source condition. In International Conference on Artificial Intelligence and Statistics, pages 3889–3897. PMLR, 2021.

[64] Ryan Rifkin and Aldebaro Klautau. In defense of one-vs-all classification. Journal of Machine Learning Research, 5:101–141, 12 2004.

[65] Mark Rudelson and Roman Vershynin. Hanson-Wright inequality and sub-Gaussian concentration. Electronic Communications in Probability, 18:1–9, 2013.

[66] Fariborz Salehi, Ehsan Abbasi, and Babak Hassibi. The impact of regularization on high-dimensional logistic regression. Advances in Neural Information Processing Systems, 32, 2019.

[67] Ohad Shamir. The implicit bias of benign overfitting. arXiv preprint arXiv:2201.11489, 2022.

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

[69] Hossein Taheri, Ramtin Pedarsani, and Christos Thrampoulidis. Sharp asymptotics and optimal performance for inference in binary models. In International Conference on Artificial Intelligence and Statistics, pages 3739–3749. PMLR, 2020.

[70] Hossein Taheri, Ramtin Pedarsani, and Christos Thrampoulidis. Fundamental limits of ridge-regularized empirical risk minimization in high dimensions. In International Conference on Artificial Intelligence and Statistics, pages 2773–2781. PMLR, 2021.

[71] Ambuj Tewari and Peter Bartlett. On the consistency of multiclass classification methods. Journal of Machine Learning Research, 8:143–157, 01 2005.

[72] Christos Thrampoulidis, Samet Oymak, and Mahdi Soltanolkotabi. Theoretical insights into multiclass classification: A high-dimensional asymptotic view. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems, volume 33, pages 8907–8920. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/6547884cea64550284728eb26b0947ef-Paper.pdf.

[73] Nilesh Tripuraneni, Ben Adlam, and Jeffrey Pennington. Overparameterization improves robustness to covariate shift in high dimensions. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, Advances in Neural Information Processing Systems, volume 34, pages 13883–13897. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/73fed7fd472e502d8908794430511f4d-Paper.pdf.

[74] Alexander Tsigler and Peter L Bartlett. Benign overfitting in ridge regression. arXiv preprint arXiv:2009.14286, 2020.

[75] Guillaume Wang, Konstantin Donhauser, and Fanny Yang. Tight bounds for minimum ℓ_1-norm interpolation of noisy data. arXiv preprint arXiv:2111.05987, 2021.

[76] Ke Wang and Christos Thrampoulidis. Benign overfitting in binary classification of Gaussian mixtures. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 4030–4034. IEEE, 2021.

[77] Ke Wang, Vidya Muthukumar, and Christos Thrampoulidis. Benign Overfitting in Multiclass Classification: All Roads Lead to Interpolation. arXiv e-prints, art. arXiv:2106.10865, June 2021.

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

[79] Jason Weston and Chris Watkins. Multi-class support vector machines. Technical report, 1998.

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

[81] Denny Wu and Ji Xu. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. Advances in Neural Information Processing Systems, 33:10112–10123, 2020.

[82] Jingfeng Wu, Difan Zou, Vladimir Braverman, and Quanquan Gu. Direction matters: On the implicit bias of stochastic gradient descent with moderate learning rate. arXiv preprint arXiv:2011.02538, 2020.

[83] Ian En-Hsu Yen, Xiangru Huang, Pradeep Ravikumar, Kai Zhong, and Inderjit Dhillon. Pd-sparse: A primal and dual sparse approach to extreme multiclass and multilabel classification. In International conference on machine learning, pages 3069–3077. PMLR, 2016.

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

[85] Tong Zhang. Statistical analysis of some multi-category large margin classification methods. Journal of Machine Learning Research, 5:1225–1251, 2004.

Checklist

1. For all authors...

   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]

   (b) Did you describe the limitations of your work? [Yes] In the conclusion as well as in the statements of the theorems.

(c) Did you discuss any potential negative societal impacts of your work? [No] Because this is a purely theoretical paper elucidating the foundations of overparameterized learning. The negative societal impacts are largely a matter of opinion whether theoretical insights from idealized models are useful. Because they are useful in teaching, we believe this is a positive social impact. However, some might believe that theoretical results for key ideas can end up having a gatekeeping effect in teaching that disadvantages populations with less access to advanced mathematics, and their mere existence therefore empowers would-be gatekeepers. However, we have attempted to mitigate this risk by building this theory together with more intuitive explanations that, while still mathematical, we believe reduce the barrier to accessing the core insights. Since this debate is orthogonal to the specific claims in this paper, we didn't feel that it needed to be addressed in the paper itself.

(d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...

   (a) Did you state the full set of assumptions of all theoretical results? [Yes]

   (b) Did you include complete proofs of all theoretical results? [Yes] But in the Supplemental material since they are too long to fit.

3. If you ran experiments...

   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A] Not really relevant since there are no empirical experiments, only plots of the regions defined by the theorems — which are themselves given by simple linear inequalities.

   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]

   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A] Not applicable since no plots are random in any way.

   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A] Not applicable since everything could be done by hand.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

   (a) If your work uses existing assets, did you cite the creators? [Yes] We cited the papers that inspired us and whose results we are building upon.

   (b) Did you mention the license of the assets? [N/A]

   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

The appendix is organized as follows. In Appendix A we provide a table of notations used throughout the paper. Appendix B provides an overall proof for Theorem 5.1 by introducing some intermediate lemmas and assuming they hold. Appendix C introduces some key tools that we need and Appendix D leverages those tools to build towards a proof of these intermediate lemmas by introducing some helper results that are needed to deal with the key challenge posed by multiclass training data. Appendix E actually proves the intermediate lemmas used in Appendix B and completes the proof. Appendix F discusses the potential looseness in our analysis and describes how we obtained Conjecture 6.1. Appendix G elaborates on the effect of fewer number of positive training examples per class in the multiclass setting and investigates an alternative setting where the total number of positive training examples per class is kept constant while we increase the number of classes. In Appendix H we provide a more detailed comparison of our work with Wang et al. [77] and Muthukumar et al. [56]. Appendix I complements the theoretical and asymptotic proofs in our paper with an empirical evaluation of relevant quantities using simulated data. Appendix J illustrates the underlying challenge provided by the regime in which regression does not generalize — namely that the empirical eigenstructure does not reveal the true nature of the underlying features. Finally, in Appendix K we provide a heuristic derivation/conjecture of our main results using an asymptotic linear-algebraic perspective.

Throughout the appendix we will assume that n is large enough for asymptotic behavior to kick in. We also will introduce various universal positive constants, indexed as c_i. These constants are all independent of n, and constants with the same index are to be treated as equal throughout this Appendix.

A Notation

We summarize the notation used in the problem setup (as well as some terms defined later) as follows:

B Proof of Theorem 5.1

We restate Theorem 5.1, our main result, here for convenience:

Theorem 5.1. ***(Asymptotic classification region in the bi-level model)****: Under the bi-level ensemble model 4.2, when the true data generating process is 1-sparse (Assumption 4.1), the probability of misclassification $P(\mathcal{E}_{err}) \to 0$ as $n \to \infty$ if the following conditions hold:*

$$t < \min\left(r, 1-r, p+1-2(q+r), p-2, 2q+r-2\right) \tag{15}$$

$$q + r > 1. \tag{16}$$

Our proof that utilizes the survival/contamination analysis framework from Muthukumar et al. [56] along with a typicality-inspired argument where the feature margin (difference between largest and second largest feature) on the test point plays a key role.

Assume without loss of generality that for the test point $\mathbf{x}_{test} \sim \mathcal{N}(0, I_d)$, the true class is α for some $\alpha \in [k]$. Let $\mathbf{x}^w_{test}$ be the weighted version of this test point. A necessary and sufficient condition for classification error is that for some $\beta \neq \alpha, \beta \in [k]$, the score associated with class β is higher than the score associated with class α. Pulling out the key terms associated with the α and β weighted features, we get:

$$\begin{aligned} \widehat{f}_\alpha[\alpha]x^w_{test}[\alpha] + \widehat{f}_\alpha[\beta]x^w_{test}[\beta] + \sum_{j \notin \{\alpha,\beta\}} \widehat{f}_\alpha[j]x^w_{test}[j] &< \widehat{f}_\beta[\alpha]x^w_{test}[\alpha] \\ &\quad + \widehat{f}_\beta[\beta]x^w_{test}[\beta] + \sum_{j \notin \{\alpha,\beta\}} \widehat{f}_\beta[j]x^w_{test}[j] \end{aligned} \tag{24}$$

$$\implies (\widehat{f}_\alpha[\alpha] - \widehat{f}_\beta[\alpha])x^w_{test}[\alpha] - (\widehat{f}_\beta[\beta] - \widehat{f}_\alpha[\beta])x^w_{test}[\beta] < \sum_{j \notin \{\alpha,\beta\}} (\widehat{f}_\beta[j] - \widehat{f}_\alpha[j])x^w_{test}[j]. \tag{25}$$

$$\tag{26}$$

Note that $\sum_{j \notin \alpha,\beta}$ refers to the sum over all feature indices 1 to d excluding α and β.

Table 1: Notation

Symbol	Definition	Dimension	Source
k	Number of classes	Scalar	Sec. 4
n	Number of training points	Scalar	Sec. 4
d	Dimension of each point — the total number of features	Scalar	Sec. 4
s	The number of favored features	Scalar	Def. 4.2
p	Parameter controlling overparameterization ($d = n^p$)	Scalar	Def. 4.2
r	Parameter controlling the number of favored features ($s = n^r$)	Scalar	Def. 4.2
a	Parameter controlling the favored weights ($a = n^{-q}$)	Scalar	Def. 4.2
t	Parameter controlling the number of classes ($k = c_k n^t$)	Scalar	Def. 4.2
c_k	The number of classes when $t = 0$ ($k = c_k n^t$)	Scalar	Def. 4.2
λ_j	Squared weight of the jth feature	Scalar	Def. 4.2
$\mathbf{x}_i$	ith training point (unweighted)	Length-n vector	Eqn. 1
ℓ_i	Class label of ith training point	Scalar	Eqn. 2
$\mathbf{x}_i^w$	ith training point (weighted)	Length-n vector	Eqn. 3
$\mathbf{X}^w$	Weighted feature matrix	$(n \times d)$-matrix	Eqn. 4
$\mathbf{z}_j$	The collected jth features of all training points	Length-n vector	Eqn. 4
$\mathbf{y}_m^{oh}$	One-hot encoding of all the training points for label m	Length-n vector	Eqn. 6
$\mathbf{Y}^{oh}$	One-hot label matrix	$(n \times k)$-matrix	Eqn. 5
$\mathbf{y}_m$	Zero-mean encoding of the training points for label m	Length-n vector	Eqn. 7
$\hat{\mathbf{f}}_m$	Learned coefficients for label m using min-norm interpolation	Length-d vector	Eqn. 10
$\mathbf{x}_{test}$	A single test point	Length-d vector	Sec. 4
$\mathbf{x}_{test}^w$	A single weighted test point	Length-d vector	Sec. 4
$\mathbf{A}$	$\mathbf{A} = \mathbf{X}^w(\mathbf{X}^w)^\top$	$(n \times n)$-matrix	Eqn. 38
$\mu_i(\mathbf{A})$	The ith eigenvalue of matrix $\mathbf{A}$, sorted in descending order	Scalar	App. B
$\mathbf{\Lambda}$	Matrix of squared feature weights: $\mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d)$	$(d \times d)$-matrix	App. B
$\widehat{h}_{\alpha,\beta}$	Relative survival $\widehat{h}_{\alpha,\beta}[j] = \lambda_j^{-1/2}(\hat{f}_\alpha[j] - \hat{f}_\beta[j])$	Length-d vector	Eqn. 28
$\mathsf{CN}_{\alpha,\beta}$	Normalizing factor $\mathsf{CN}_{\alpha,\beta} = \sqrt{\left(\sum_{j\notin\{\alpha,\beta\}} \lambda_j^2 (\widehat{h}_{\beta,\alpha}[j])^2\right)}$	Scalar	Eqn. 34
$\lVert\cdot\rVert_{\psi_2}$	The sub-Gaussian norm of a scalar random variable	Scalar	Eqn. 74
$\bar{\mu}$	Center of the eigenvalue bounds for $\mathbf{A}^{-1}$, $\bar{\mu} = \frac{1}{\sum_j \lambda_j}$	Scalar	Eqn. 41
$\Diamond$	Deviation term in eigenvalue bounds for $\mathbf{A}$	Scalar	Eqn. 77
Δ_μ	Deviation term in eigenvalue bounds for $\mathbf{A}^{-1}$	Scalar	Eqn. 43

By converting into the unweighted feature space we obtain,

$$\lambda_\alpha \widehat{h}_{\alpha,\beta}[\alpha] x_{test}[\alpha] - \lambda_\beta \widehat{h}_{\beta,\alpha}[\beta] x_{test}[\beta] < \sum_{j\notin\{\alpha,\beta\}} \lambda_j \widehat{h}_{\beta,\alpha}[j] x_{test}[j], \tag{27}$$

where we introduce the short-hand notation,

$$\widehat{h}_{\alpha,\beta}[j] = \lambda_j^{-1/2}(\hat{f}_\alpha[j] - \hat{f}_\beta[j]) \tag{28}$$

$$\widehat{h}_{\beta,\alpha}[j] = \lambda_j^{-1/2}(\hat{f}_\beta[j] - \hat{f}_\alpha[j]). \tag{29}$$

Since both α and β are favored feature indices, by leveraging the definition of the bi-level model and denoting $\lambda_\alpha = \lambda_\beta = \lambda$, we get

$$\lambda \left(\widehat{h}_{\alpha,\beta}[\alpha] x_{test}[\alpha] - \widehat{h}_{\beta,\alpha}[\beta] x_{test}[\beta] \right) < \sum_{j\notin\{\alpha,\beta\}} \lambda_j \widehat{h}_{\beta,\alpha}[j] x_{test}[j]. \tag{30}$$

Next, we perform some algebraic manipulations,

$$\lambda\left(\widehat{h}_{\alpha,\beta}[\alpha]x_{test}[\alpha] - \widehat{h}_{\beta,\alpha}[\beta]x_{test}[\beta]\right) < \sum_{j\notin\{\alpha,\beta\}} \lambda_j \widehat{h}_{\beta,\alpha}[j]x_{test}[j] \tag{31}$$

$$\implies \lambda\widehat{h}_{\alpha,\beta}[\alpha](x_{test}[\alpha] - x_{test}[\beta]) + \lambda x_{test}[\beta](\widehat{h}_{\alpha,\beta}[\alpha] - \widehat{h}_{\beta,\alpha}[\beta]) < \sum_{j\notin\{\alpha,\beta\}} \lambda_j \widehat{h}_{\beta,\alpha}[j]x_{test}[j] \tag{32}$$

$$\implies \lambda\widehat{h}_{\alpha,\beta}[\alpha]\left((x_{test}[\alpha] - x_{test}[\beta]) + x_{test}[\beta]\frac{\widehat{h}_{\alpha,\beta}[\alpha] - \widehat{h}_{\beta,\alpha}[\beta]}{\widehat{h}_{\alpha,\beta}[\alpha]}\right) < \sum_{j\notin\{\alpha,\beta\}} \lambda_j \widehat{h}_{\beta,\alpha}[j]x_{test}[j]. \tag{33}$$

We divide both sides by the quantity $\mathsf{CN}_{\alpha,\beta}$ defined as,

$$\mathsf{CN}_{\alpha,\beta} = \sqrt{\left(\sum_{j\notin\{\alpha,\beta\}} \lambda_j^2 (\widehat{h}_{\beta,\alpha}[j])^2\right)}. \tag{34}$$

This normalizes the RHS of (33) to have a standard normal distribution. Thus, the necessary and sufficient condition for a misclassification error is for some $\beta \neq \alpha, \beta \in [k]$,

$$\frac{\lambda\widehat{h}_{\alpha,\beta}[\alpha]}{\mathsf{CN}_{\alpha,\beta}}\left((x_{test}[\alpha] - x_{test}[\beta]) + x_{test}[\beta]\frac{\widehat{h}_{\alpha,\beta}[\alpha] - \widehat{h}_{\beta,\alpha}[\beta]}{\widehat{h}_{\alpha,\beta}[\alpha]}\right) < \frac{1}{\mathsf{CN}_{\alpha,\beta}} \sum_{j\notin\{\alpha,\beta\}} \lambda_j \widehat{h}_{\beta,\alpha}[j]x_{test}[j]. \tag{35}$$

A sufficient condition for correct classification can then be obtained by ensuring that the smallest potential value of the LHS is still greater than the value of the RHS for all values of β. Thus, we obtain a sufficient condition for correct classification by appropriately minimizing or maximizing quantities over competing feature indices $\beta \neq \alpha, \beta \in [k]$ (for notational convenience we simply denote this as $\min_\beta$ or $\max_\beta$).

$$\underbrace{\frac{\min_\beta \lambda\widehat{h}_{\alpha,\beta}[\alpha]}{\max_\beta \mathsf{CN}_{\alpha,\beta}}}_{\text{SU/CN ratio}}\left(\underbrace{\min_\beta (x_{test}[\alpha] - x_{test}[\beta])}_{\text{closest feature margin}} - \underbrace{\max_\beta |x_{test}[\beta]|}_{\text{largest competing feature}} \underbrace{\cdot \max_\beta \left|\frac{\widehat{h}_{\alpha,\beta}[\alpha] - \widehat{h}_{\beta,\alpha}[\beta]}{\widehat{h}_{\alpha,\beta}[\alpha]}\right|}_{\text{survival variation}}\right)$$
$$> \underbrace{\max_\beta \frac{1}{\mathsf{CN}_{\alpha,\beta}}\left(\sum_{j\notin\{\alpha,\beta\}} \lambda_j \widehat{h}_{\beta,\alpha}[j]x_{test}[j]\right)}_{\text{normalized contamination}}. \tag{36}$$

We will show that under the conditions specified in Theorem 5.1, with sufficiently high probability, the relevant survival to contamination SU/CN ratio grows at a polynomial rate n^v for some $v > 0$, the closest feature margin shrinks at a less-than-polynomial rate $1/\sqrt{\ln nk}$, and the survival variation decays at a polynomial rate n^{-u} for some $u > 0$. Further, the magnitudes of the largest competing feature and the normalized contamination are no more than $2\sqrt{\ln(nk)}$. Here, we leverage the idea of typicality-style proofs in information theory [17] to avoid unnecessarily loose union bounds that end up being dominated by the atypical behavior of quantities. In our case, by pulling the feature margin out explicitly, we can just deal with its typical behavior. Similarly, the typical behavior of the largest competing feature and the true feature is all that matters. Before we proceed with the rest of our proof we remind the reader of a few important definitions.

Recall from (10) that our learned feature coefficients are

$$\hat{\mathbf{f}}_m = (\mathbf{X}^w)^\top \left(\mathbf{X}^w(\mathbf{X}^w)^\top\right)^{-1}\mathbf{y}_m. \tag{37}$$

Let

$$\mathbf{A} = \mathbf{X}^w (\mathbf{X}^w)^\top. \tag{38}$$

Then we can express our learned coefficients as

$$\hat{f}_m[j] = \sqrt{\lambda_j} \mathbf{z}_j^\top \mathbf{A}^{-1} \mathbf{y}_m, \tag{39}$$

where $\mathbf{z}_j \in \mathbb{R}^n$ contains the j^{th} features of all n training points. The rows of $\mathbf{X}^w$ are i.i.d. Gaussians with covariance matrix $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d)$. Let $\mu_1(\mathbf{A})$ denote the largest eigenvalue and $\mu_n(\mathbf{A})$ denote the smallest eigenvalue of $\mathbf{A}$ respectively, with $\mu_i(\mathbf{A})$ being the i-th largest eigenvalue of $\mathbf{A}$.

Next, we state a useful lemma adapted from Bartlett et al. [4] [12] that bounds the eigenvalues of $\mathbf{A}^{-1}$. Subsequent lemmas will utilize these eigenvalue bounds.

Lemma B.1. ***(Eigenvalue bounds on*** $\mathbf{A}^{-1}$ ***adapted from Bartlett et al. [4])****:*
If $\mathbf{\Lambda}$ is such that $\lozenge \ll \sum_j \lambda_j$, then with probability at least $(1 - 2e^{-n})$,

$$\bar{\mu} - \Delta_\mu \leq \mu_n(\mathbf{A}^{-1}) \leq \mu_1(\mathbf{A}^{-1}) \leq \bar{\mu} + \Delta_\mu, \tag{40}$$

where,

$$\bar{\mu} = \frac{1}{\sum_j \lambda_j} \tag{41}$$

$$\lozenge = \frac{32}{9}\left(\lambda_1(1 + \ln 9)n + \sqrt{(1 + \ln 9)n \sum_j \lambda_j^2}\right) \tag{42}$$

$$\Delta_\mu = \bar{\mu}\left(\frac{\lozenge}{\sum_j \lambda j} + \Theta\left(\frac{\lozenge}{\sum_j \lambda j}\right)^2\right). \tag{43}$$

Further this implies that with probability at least $(1 - 2e^{-n})$,

$$\left|\mu_i(\mathbf{A}^{-1} - \bar{\mu}\mathbf{I}_n)\right| \leq \Delta_\mu \tag{44}$$

for all $i \in [n]$.

The subsequent lemmas bound the feature margin, survival, contamination and survival variation terms, utilizing tools from [4] and building on results from [56].

Lemma B.2. ***(Lower bound on the closest feature margin as*** $k \to \infty$***)****: For any constant $\varepsilon > 0$, there exists a constant θ such that, for sufficiently large k with probability at least $(1 - \varepsilon)$,*

$$\min_{\beta: 1 \leq \beta \neq \alpha \leq k} (x_{test}[\alpha] - x_{test}[\beta]) \geq \frac{\theta}{\sqrt{2\ln(k)}}. \tag{45}$$

Here, α is fixed and corresponds to the index of the true class — i.e. α corresponds to the index of the maximum feature among the first k features.

Lemma B.3. ***(Lower bound on the closest feature margin when*** k ***is constant)****: If $k = c_k$ for some fixed constant c_k, for any constant $\varepsilon > 0$, there exists a constant $\varepsilon' > 0$ such that*

$$\Pr\left(\min_{\beta,\gamma: 1 \leq \beta \neq \gamma \leq c_k} |x_{test}[\beta] - x_{test}[\gamma]| \geq \varepsilon'\right) \geq 1 - \varepsilon. \tag{46}$$

Thus, with probability at least $(1 - \varepsilon)$,

$$\min_{\beta: 1 \leq \beta \neq \alpha \leq k} (x_{test}[\alpha] - x_{test}[\beta]) \geq \varepsilon'. \tag{47}$$

Here, α is fixed and corresponds to the index of the true class — i.e. α corresponds to the index of the maximum feature among the first k features.

[12] More precisely this lemma appeared in the first version of this work at `https://arxiv.org/pdf/1906.11300v1.pdf`. In subsequent versions the authors use a slightly weaker version of this result since it is sufficient for their purposes.

Lemma B.4. *(**Lower bound on relative survival of true feature**): For any fixed $\beta \in [k]$, $\beta \neq \alpha$, with $\lambda_\alpha = \lambda_\beta = \lambda$ we have with probability at least $(1 - 5/(nk))$,*

$$\lambda \widehat{h}_{\alpha,\beta}[\alpha] \geq \lambda \left(c_{10} \bar{\mu} \frac{n}{k} \sqrt{\ln(k)} - c_9 (\bar{\mu} \sqrt{n} \sqrt{\ln(nk)} + \Delta_\mu \cdot n/\sqrt{k}) \right), \tag{48}$$

for universal positive constants c_9 and c_{10}.

By substituting the asymptotic behavior of parameters from our bi-level ensemble model we get the following corollary:

Corollary B.4.1. *Under the bi-level ensemble model 4.2, for any fixed $\beta \in [k]$, $\beta \neq \alpha$, $\lambda_\alpha = \lambda_\beta = \lambda$ if $t < 1/2$, $t < 2(q + r - 1)$ and $1 < q + r < (p + 1)/2$, with probability at least $(1 - 5/(nk))$,*

$$\lambda \widehat{h}_{\alpha,\beta}[\alpha] \geq c_{12} n^{1-q-r-t} \sqrt{\ln(k)}, \tag{49}$$

for universal positive constant c_{12}.

Lemma B.5. *(**Upper bound on contamination**): For any fixed $\beta \in [k]$, $\beta \neq \alpha$, with probability at least $(1 - 7/(nk))$,*

$$\mathsf{CN}_{\alpha,\beta} \leq c_7 (\bar{\mu} \sqrt{\frac{n}{k}} \cdot \sqrt{\ln(ndk)} + \Delta_\mu \cdot n/\sqrt{k}) \sqrt{\sum \lambda_j^2}, \tag{50}$$

for universal positive constant c_7.

As before, for our bi-level ensemble model we have the corollary:

Corollary B.5.1. *Under the bi-level model 4.2, in the regime $1 < q + r < (p + 1)/2$, with probability at least $(1 - 7/(nk))$,*

$$\mathsf{CN}_{\alpha,\beta} \leq c_{13} n^{(1-t-p)/2 + \max(0, 3/2 - q - r) + \max(0, p/2 - q - r/2)} \sqrt{\ln(ndk)}, \tag{51}$$

for universal positive constant c_{13}.

Lemma B.6. *(**Upper bound on survival variance**): For any fixed competing feature $\beta \in [k]$, $\beta \neq \alpha$ with $\lambda_\alpha = \lambda_\beta$, we have with probability at least $(1 - 15/(nk))$,*

$$\frac{\widehat{h}_{\alpha,\beta}[\alpha] - \widehat{h}_{\beta,\alpha}[\beta]}{\widehat{h}_{\alpha,\beta}[\alpha]} \leq \frac{2c_9 (\bar{\mu} \sqrt{n} \sqrt{\ln(nk)} + \Delta_\mu \cdot n/\sqrt{k})}{c_{10} \bar{\mu} \frac{n}{k} \sqrt{\ln(k)} - c_9 (\bar{\mu} \sqrt{n} \sqrt{\ln(nk)} + \Delta_\mu \cdot n/\sqrt{k})}, \tag{52}$$

for universal positive constants c_9 and c_{10}.

As before, we can also obtain the asymptotic bound:

Corollary B.6.1. *Under the bi-level ensemble model 4.2, for any fixed $\beta \in [k]$, $\beta \neq \alpha$, if $t < 1/2$, $t < 2(q + r - 1)$, and $1 < q + r < (p + 1)/2$, with probability at least $(1 - 15/(nk))$,*

$$\frac{\widehat{h}_{\alpha,\beta}[\alpha] - \widehat{h}_{\beta,\alpha}[\beta]}{\widehat{h}_{\alpha,\beta}[\alpha]} < n^{-u}, \tag{53}$$

for large enough n for some fixed $u > 0$.

Next, we assume that the lemmas and corollaries stated above are true and complete the proof for Theorem 5.1. We provide proofs for these lemmas in Appendices C, D and E.

Assume we are in the regime where $t < 1/2$, $t < 2(q + r - 1)$, and $1 < q + r < (p + 1)/2$, so all our corollaries above hold. Denote the misclassification event as $\mathcal{E}_{err}$ and let $\varepsilon > 0$ be an arbitrarily chosen constant.

Substitute Corollaries B.4.1, B.5.1, and B.6.1 into (22), applying them on all $1 \leq \beta \neq \alpha \leq k$. They hold with probability at least $1 - 5/(nk)$, $1 - 7/(nk)$, and $1 - 15/(nk)$ respectively for a given test point and choice of β. So by the union bound across the three bounds and all $k - 1$ choices of β, with probability at most $27/n$, one of these corollaries will not hold for our test point for some β. Let this failure event be denoted E_1.

In the case when E_1 does not occur, misclassification occurs only if

$$\frac{c_{12} \sqrt{\ln(k)}}{c_7 \sqrt{\ln(ndk)}} n^v \left(\min_\beta (x_{test}[\alpha] - x_{test}[\beta]) - \max_\beta |x_{test}[\beta]| \cdot n^{-u} \right) < \max_\beta Z^{(\beta)}, \tag{54}$$

where we define the exponent

$$v = 1 - q - r - t - (1-t-p)/2 - \max\left(0, \frac{3}{2} - q - r\right) - \max\left(0, \frac{p}{2} - q - \frac{q}{2}\right) \tag{55}$$

$$= \frac{p+1}{2} - q - r - \frac{t}{2} - \max\left(0, \frac{3}{2} - q - r, \frac{p}{2} - q - \frac{r}{2}, \frac{3}{2} - 2q - \frac{3r}{2}\right), \tag{56}$$

and

$$Z^{(\beta)} = \frac{1}{\mathsf{CN}_{\alpha,\beta}} \left(\sum_{j \notin \{\alpha,\beta\}} \lambda_j \widehat{h}_{\beta,\alpha}[j] x_{test}[j] \right). \tag{57}$$

For each class β, observe that we have $Z^{(\beta)} \sim \mathcal{N}(0,1)$.[13] Thus, by the Gaussian tail bound, for each β with probability at least $(1 - 1/(nk))$,

$$Z^{(\beta)} < \sqrt{2\ln(nk)}. \tag{58}$$

So by the union bound over all k classes β, with probability at least $(1 - 1/n)$,

$$\max_{\beta} Z^{(\beta)} < \sqrt{2\ln(nk)}. \tag{59}$$

Let the failure event where this is not the case be E_2.

An identical argument shows that with probability at least $(1 - 2/n)$, $\max_\beta |x_{test}[\beta]| \leq \sqrt{2\ln(nk)}$. Let E_3 be the failure event where this is not the case.

From Lemma B.2, we know with probability $1 - \varepsilon$ that, if $t > 0$, then for sufficiently large n (and so sufficiently large k)

$$\min_{\beta} \left(x_{test}[\alpha] - x_{test}[\beta]\right) > \frac{\theta}{\sqrt{2\ln(k)}}. \tag{60}$$

If $t = 0$ and $k = c_k$, then Lemma B.3 states that, with probability $1 - \varepsilon$,

$$\Pr\left(\min_{1 \leq \beta \neq \gamma \leq c_k} |x_{test}[\beta] - x_{test}[\gamma]| \geq \varepsilon' \right) \geq 1 - \varepsilon, \tag{61}$$

for some constant ε'. Let the ε-probability event of the appropriate margin bound (depending on whether $t = 0$ or $t > 0$) being violated be the error event E_4.

Assuming E_1, E_2, E_3, and E_4 all do not take place, misclassification can only occur if

$$\frac{c_{12}\sqrt{\ln(k)}}{c_7\sqrt{\ln(ndk)}} n^v \left(\min\left(1 - \varepsilon, \frac{\theta}{\sqrt{2\ln(k)}}\right) - \sqrt{2\ln(nk)} n^{-u} \right) < \sqrt{2\ln(nk)}. \tag{62}$$

Clearly, if $v > 0$, then (for sufficiently large n) misclassification becomes asymptotically impossible (except via the specified error events), since the LHS of the above grows asymptotically faster than the RHS.

The union bound shows that the probability of any of E_1, E_2, E_3, E_4 occurring tends to ε as $n \to \infty$ (since the probability of the first three tend to zero). So in the regime where

$$t < \frac{1}{2} \tag{63}$$

$$t < 2(q + r - 1) \tag{64}$$

$$q + r > 1 \tag{65}$$

$$\frac{p+1}{2} > q + r + \frac{t}{2} + \max\left(0, \frac{3}{2} - q - r\right) + \max\left(0, \frac{p}{2} - q - \frac{r}{2}\right), \tag{66}$$

[13] To be precise, here we can think of fixing the training data and looking purely at the randomness arising from the features in the test point. The resulting $Z^{(\beta)}$ is a standard normal. Since we are using the union bound in our proof finally, this is sufficient for our purposes.

the probability of misclassification tends to ε for sufficiently large n, for any $\varepsilon > 0$.

Consolidation of the above bounds produces the conditions [14]

$$t < \min\left(1 - r, p + 1 - 2(q + r), p - 2, 2q + r - 2\right) \tag{71}$$

$$q + r > 1. \tag{72}$$

Finally, note that the condition $t < r$ comes from the definition of the bi-level model (4.2). This condition simply states that for good generalization we must favor all the features used to determine classes. Since the analysis above holds for any ε, we see that within this regime the probability of misclassification must approach zero in the limit. This completes the proof. Note that while we show that probability of misclassification goes to zero, we do not show it to do so at any particular rate, because the result from Lemma B.2 does not specify the rate of convergence.

C Useful results from elsewhere that we need

This section collects results that are used in our proof, but which come from elsewhere or are lightly adapted to our purposes.

Hanson-Wright inequality [65]: Let $\mathbf{z}$ be a random vector composed of i.i.d. random variables that are zero mean and with sub-Gaussian norm at most K. The sub-Gaussian norm $\|\xi\|_{\psi_2}$ of a random variable ξ is defined as in Rudelson and Vershynin [65],

$$\|\xi\|_{\psi_2} = \inf_{K>0} K \tag{73}$$

$$\text{s.t. } \mathbb{E}\exp\left(\xi^2/K^2\right) \leq 2. \tag{74}$$

Then, there exists universal constant $c > 0$ such that for any positive semi-definite matrix M and for every $t \geq 0$, we have

$$\Pr\left[|\mathbf{z}^T M\mathbf{z} - \mathbb{E}[\mathbf{z}^T M\mathbf{z}]| > t\right] \leq 2\exp\left\{-c\min\left\{\frac{t^2}{K^4||M||_{\mathsf{F}}^2}, \frac{t}{K^2||M||_{\mathsf{op}}}\right\}\right\} \tag{75}$$

The next result bounds the eigenvalues of the $n \times n$ matrix $\mathbf{A} = \mathbf{X}^w(\mathbf{X}^w)^\top$, where recall that the rows of $\mathbf{X}^w$ are i.i.d. Gaussians with covariance matrix $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d)$. Let $\mu_1(\mathbf{A})$ denote the largest eigenvalue and $\mu_n(\mathbf{A})$ denote the smallest eigenvalue of $\mathbf{A}$ respectively.

From Bartlett et al. [4] , we have the following result

Lemma C.1. *With probability at least $(1 - 2e^{-n})$, the eigenvalues of $\mathbf{A}$ satisfy:*

$$\sum_j \lambda_j - \Diamond \leq \mu_n(\mathbf{A}) \leq \mu_1(\mathbf{A}) \leq \sum_j \lambda_j + \Diamond, \tag{76}$$

where,

$$\Diamond = \frac{32}{9}\left(\lambda_1(1 + \ln 9)n + \sqrt{(1 + \ln 9)n\sum_j \lambda_j^2}\right). \tag{77}$$

[14] We can simplify (66) as follows:

$$\frac{p+1}{2} > q + r + \frac{t}{2} \implies t < p + 1 - 2(q + r) \tag{67}$$

$$\frac{p+1}{2} > q + r + \frac{t}{2} + \frac{3}{2} - q - r \implies t < p - 2 \tag{68}$$

$$\frac{p+1}{2} > q + r + \frac{t}{2} + \frac{p}{2} - q - \frac{r}{2} \implies t < 1 - r \tag{69}$$

$$\frac{p+1}{2} > q + r + \frac{t}{2} + \frac{3}{2} - q - r + \frac{p}{2} - q - \frac{r}{2} \implies t < 2q + r - 2. \tag{70}$$

Then we note that $t < \min(r, 1 - r) \implies t < 1/2$.

Next, as stated previously in Lemma B.1 we will use this result to obtain bounds on the eigenvalues of $\mathbf{A}^{-1}$ assuming that $\mathbf{\Lambda}$ is such that $\Diamond \ll \sum_j \lambda_j$.[15]

Lemma B.1. ***(Eigenvalue bounds on*** $\mathbf{A}^{-1}$ ***adapted from Bartlett et al. [4]):*** *If* $\mathbf{\Lambda}$ *is such that* $\Diamond \ll \sum_j \lambda_j$*, then with probability at least* $(1-2e^{-n})$*,*

$$\bar{\mu} - \Delta_\mu \le \mu_n(\mathbf{A}^{-1}) \le \mu_1(\mathbf{A}^{-1}) \le \bar{\mu} + \Delta_\mu, \tag{40}$$

where,

$$\bar{\mu} = \frac{1}{\sum_j \lambda_j} \tag{41}$$

$$\Diamond = \frac{32}{9}\left(\lambda_1(1+\ln 9)n + \sqrt{(1+\ln 9)n\sum_j \lambda_j^2}\right) \tag{42}$$

$$\Delta_\mu = \bar{\mu}\left(\frac{\Diamond}{\sum_j \lambda j} + \Theta\left(\frac{\Diamond}{\sum_j \lambda j}\right)^2\right). \tag{43}$$

Further this implies that with probability at least $(1-2e^{-n})$*,*

$$\left|\mu_i(\mathbf{A}^{-1} - \bar{\mu}\mathbf{I}_n)\right| \le \Delta_\mu \tag{44}$$

for all $i \in [n]$*.*

Proof. Let $S = \sum_j \lambda_j$.

$$\frac{1}{S+\Diamond} = \frac{1}{S}\left(1+\frac{\Diamond}{S}\right)^{-1} \tag{78}$$

$$= \frac{1}{S}\left(1-\frac{\Diamond}{S}+\Theta\left(\frac{\Diamond}{S}\right)^2\right) \tag{79}$$

$$= \bar{\mu} - \Delta_\mu, \tag{80}$$

and analogously $(S-\Diamond)^{-1} = \bar{\mu} + \Delta_\mu$. Taking reciprocals of everything in the inequality 76, and since the eigenvalues of $\mathbf{A}$ and $\mathbf{A}^{-1}$ are reciprocals of each other, the desired result follows.

□

As a Corollary of Lemma B.1:

Corollary C.1.1. ***(Asymptotic eigenvalue bounds on*** $\mathbf{A}^{-1}$***)*** *Considering the asymptotic scaling of the model parameters from the bi-level model (Definition 4.2), in the regime* $1 < q + r < (1+p)/2$*,*

$$\bar{\mu} = n^{-p} \tag{81}$$

$$\Delta_\mu \le c_4 n^{1-p-q-r} \ll \bar{\mu}, \tag{82}$$

where $\bar{\mu}$ *and* Δ_μ *are defined as in Lemma B.1, and* c_4 *is a universal constant.*

Proof. From the asymptotic scaling of the λ_j from (13) and (14), we see that (from the definition provided in Lemma B.1)

$$\bar{\mu} = \frac{1}{\sum_j \lambda_j} \tag{83}$$

$$= \frac{1}{n^r n^{p-q-r} + (n^p - n^r)(1-n^q)\cdot n^p/(n^p - n^r))} \tag{84}$$

$$= \frac{1}{n^{p-q} + n^p - n^{p-q}} \tag{85}$$

$$= n^{-p}. \tag{86}$$

[15] Note that in the regime $q + r < 1$ (where regression works [56]), we do not have $\Diamond \ll \lambda_j$ and in such scenarios we cannot simply rely on eigenvalue bounds and need to use other techniques in the proof.

Next, we have that

$$\begin{aligned}
\Diamond &= \frac{32}{9}\left(\lambda_1(1+\ln 9)n + \sqrt{(1+\ln 9)n\sum_j \lambda_j^2}\right) && (87)\\
&\le c_1 n^{1+p-q-r} + c_2\sqrt{n(n^r n^{2p-2q-2r} + (n^p - n^r))} && (88)\\
&\le c_1 n^{1+p-q-r} + c_2\sqrt{n^{1+2p-2q-r} + n^{1+p}} && (89)
\end{aligned}$$

for constants c_1 and c_2,

The second term is of the order $n^{\max((1-r)/2+p-q,(1+p)/2)}$. Thus, in the regime $q + r < (1 + p)/2$, and since $r < 1$ we have $1 + p - q - r > (1 - r)/2 + p - q$ and $1 + p - q - r > (1 + p)/2$ and the first term dominates.

Thus, $\Diamond \le c_3 n^{1+p-q-r}$ for some constant c_3 and sufficiently large n.

Observe that since $q + r > 1$, $\Diamond \ll \sum_j \lambda_j = n^p$. Thus, we can substitute into our relation for Δ_μ from Lemma B.1, to see that

$$\begin{aligned}
\Delta_\mu &= \bar{\mu}\left(\frac{\Diamond}{\sum_j \lambda j} + \Theta\left(\frac{\Diamond}{\sum_j \lambda j}\right)^2\right) && (90)\\
&\le n^{-p}\left((c_3 n^{1+p-q-r})(n^{-p}) + \Theta((c_3 n^{1+p-q-r})^2(n^{-p})^2)\right) && (91)\\
&= n^{-p}(c_3 n^{1-q-r} + \Theta(c_3 n^{2(1-q-r)})). && (92)
\end{aligned}$$

In the regime where $q + r > 1$, the first term in the sum dominates the second, giving us,

$$\Delta_\mu \le c_4 n^{1-p-q-r} \tag{93}$$

for some constant c_4 and sufficiently large n. This completes the proof. □

Finally, in this section, we restate well-known bounds concerning Gaussian random variables.

Lemma C.2. *Chi-squared tail bound:*
Let $\mathbf{z} \sim \mathcal{N}(0, I_n)$. *For any* $\delta \in (0, 1)$, *with probability at least* $(1 - 2e^{-n\delta^2})$ *we have:*

$$n(1-\delta) \le \|\mathbf{z}\|^2 \le n(1+\delta). \tag{94}$$

From bounds on the expectation of the maximum of k Gaussians:

Lemma C.3. *Let* $\mathbf{z}_\alpha = \max_{1\le j\le k} \mathbf{z}_j$ *where* $\mathbf{z}_j \sim \mathcal{N}(0, 1)$. *Then,*

$$\frac{1}{\sqrt{\pi \ln 2}} \cdot \sqrt{\ln k} \le \mathbb{E}[\mathbf{z}_\alpha] \le \sqrt{2} \cdot \sqrt{\ln k}. \tag{95}$$

D Utility Bounds

The big technical challenge in moving from binary classification (as studied in Muthukumar et al. [56]) to multiclass classification has to do with the nature of the training data. Whereas for binary classification one could change coordinates so that the binary labels only depended on a single Gaussian random variable and were independent of all other directions of Gaussian variation in the covariates, no such change of coordinates exists for multiclass labels. The one-hot-style encoding of the labels fundamentally depends on the realizations of all k of the Gaussian random variables representing each of the k classes. This means that we can no longer simply leverage independence to simplify the analysis and certain clever approaches used to invoke Hanson-Wright are no longer available to us. However, the need remains to appropriately bound quadratic forms of the form $\left|\mathbf{z}_j^\top \mathbf{A}^{-1}\boldsymbol{\Delta} y\right|$ both for the cases when j represents a feature that is not dominant in the computation of $\boldsymbol{\Delta} y$ as well as in cases where j represents a feature that is dominant in $\boldsymbol{\Delta} y$. To be able to control such quantities in the absence of the independence we could leverage in the binary case, this section of the Appendix derives two lemmas which can be viewed as helper bounds. These bounds will later be used to bound the various quantities from (22). Because our focus is on the asymptotic scaling, we will use c_i to denote the appropriate global constants.

In the subsequent lemmas, $\bar{\mu}$ and Δ_μ are defined as in the bounds on the eigenvalues of $\mathbf{A}^{-1}$ from Lemma B.1.

The following lemma is used to upper-bound the contamination term $\mathsf{CN}_{\alpha,\beta}$ in Lemma B.5:

Lemma D.1. *Let $\mathbf{\Delta}y = \mathbf{y}_\alpha - \mathbf{y}_\beta$. Let α, β, and j be distinct. Then, with probability at least $(1 - 7/(ndk))$, we have,*

$$\left|\mathbf{z}_j^\top \mathbf{A}^{-1}\mathbf{\Delta}y\right| \le c_7(\bar{\mu}\sqrt{\frac{n}{k}} \cdot \sqrt{\ln(ndk)} + \Delta_\mu \cdot n/\sqrt{k}), \tag{96}$$

for some constant c_7.

This next lemma is used to bound the numerator of the survival variation term from (22):

Lemma D.2. *Let $\mathbf{\Delta}y = \mathbf{y}_\alpha - \mathbf{y}_\beta$. With probability at least $(1 - 5/(nk))$, we have each of*

$$\mathbf{z}_\alpha^\top \mathbf{A}^{-1}\mathbf{\Delta}y \le \bar{\mu}(\mathbb{E}[\mathbf{z}_\alpha^\top \mathbf{y}_\alpha] - \mathbb{E}[\mathbf{z}_\alpha^\top \mathbf{y}_\beta]) + c_9(\bar{\mu}\sqrt{n}\sqrt{\ln(nk)} + \Delta_\mu \cdot n/\sqrt{k}) \tag{97}$$

$$\mathbf{z}_\alpha^\top \mathbf{A}^{-1}\mathbf{\Delta}y \ge \bar{\mu}(\mathbb{E}[\mathbf{z}_\alpha^\top \mathbf{y}_\alpha] - \mathbb{E}[\mathbf{z}_\alpha^\top \mathbf{y}_\beta]) - c_9(\bar{\mu}\sqrt{n}\sqrt{\ln(nk)} + \Delta_\mu \cdot n/\sqrt{k}), \tag{98}$$

for some constant c_9.

The following corollary of the above is used to lower-bound the relative survival $\widehat{h}_{\alpha,\beta}[\alpha]$, which in turn bounds the SU/CN ratio and the denominator of the survival variation term:

Corollary D.2.1. *Let $\mathbf{\Delta}y = \mathbf{y}_\alpha - \mathbf{y}_\beta$. With probability at least $(1 - 5/(nk))$, we have,*

$$\mathbf{z}_\alpha^\top \mathbf{A}^{-1}\mathbf{\Delta}y \ge c_{10}\bar{\mu}\frac{n}{k}\sqrt{\ln(k)} - c_9(\bar{\mu}\sqrt{n}\sqrt{\ln(nk)} + \Delta_\mu \cdot n/\sqrt{k}), \tag{99}$$

for some constant c_{10}.

D.1 Proof of Lemma D.1

We will write $\mathbf{A}^{-1} = \bar{\mu}\mathbf{I}_n + \mathbf{\Delta}A_{inv}$, and split up the expression $\mathbf{z}_j^\top \mathbf{A}^{-1}\mathbf{\Delta}y$ into components involving $\bar{\mu}\mathbf{I}_n$, and components involving $\mathbf{\Delta}A_{inv}$. To bound the first term, we will use Hanson-Wright, and to bound the second we will use Cauchy-Schwartz. Throughout the proof, we rely on the concentration of the eigenvalues of $\mathbf{A}^{-1}$.

Next, we bound the first term (we set aside the constant $\bar{\mu}$ for now and deal with it later).

D.1.1 Bounds on $\mathbf{z}_j^T(\mathbf{y}_\alpha - \mathbf{y}_\beta)$

Throughout this section, let j be a feature index distinct from α and β. Define the diagonal matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$ with diagonal entries given by:

$$M_{ii} = \begin{cases} 1, & \text{if } \Delta y[i] \neq 0 \\ 0, & \text{otherwise} \end{cases}. \tag{100}$$

In other words, M_{ii} is 1 only if training point i belongs to class α or β and is 0 otherwise. Thus for each $i \in [n]$, $M_{ii} \sim \mathit{Bernoulli}(2/k)$ and are independent of each other. We introduce this matrix $\mathbf{M}$ to ensure that our bound reflects the fact that most of the entries of $\mathbf{\Delta}y$ are 0. In particular $\Delta y[i] \neq 0$ only if point i belongs to class α or β and only contains roughly $2n/k$ non-zero entries.[16] Note that we have by definition,

$$\mathbf{z}_j^T\mathbf{\Delta}y = \mathbf{z}_j^T\mathbf{M}\mathbf{\Delta}y. \tag{101}$$

Our strategy is to bound $\mathbf{z}_j^\top \mathbf{M}\mathbf{\Delta}y$ for every typical realization $\mathcal{M}$ of the random variable $\mathbf{M}$ using the Hanson-Wright inequality. Subsequently, we will apply these bounds with high probability over typical realizations of $\mathbf{M}$ that satisfy the Proposition below, which merely asserts that with high probability, the number of 1s in $\mathbf{\Delta}y$ is close to its expected value.

Proposition D.1. *For $\delta \in (0, 1)$, with probability at least $(1 - 2e^{-\frac{2n\delta^2}{3k}})$, the trace of $\mathbf{M}$ is bounded as:*

$$(1-\delta)\frac{2n}{k} \le \|\mathbf{\Delta}y\|_2^2 = \operatorname{Tr}(\mathbf{M}) \le (1+\delta)\frac{2n}{k}. \tag{102}$$

[16] An alternative bounding technique that first converted $\mathbf{z}_j^\top\mathbf{\Delta}y$ to a quadratic form and applied Hanson-Wright would be looser by a factor of $\sqrt{k}$ if we did not introduce $\mathbf{M}$.

Proof. Note that $\mathrm{Tr}(\mathbf{M})$ is the sum of n i.i.d Bernoulli random variables with mean $2/k$. The result follows by application of the Chernoff bound. □

Note that once we fix the realization $\mathcal{M}$, the distributions of $\mathbf{z}_j$ and $\boldsymbol{\Delta} y$ will now have to be conditioned on this realization and we need to deal with the modified distributions while applying the Hanson-Wright inequality. In particular, once we know that a feature was not the winning feature, it is no longer zero-mean.

Now,

$$\begin{aligned}
\mathbf{z}_j^T \mathcal{M} \boldsymbol{\Delta} y &= \sum_i z_j[i] \mathcal{M}_{ii} \Delta y[i] && (103)\\
&= \sum_{i:\mathcal{M}_{ii}=1} z_j[i] \Delta y[i] && (104)\\
&= \sum_{i:\mathcal{M}_{ii}=1} \left(z_j[i] - \mathbb{E}[z_j[i] \mid M_{ii}=1]\right) \Delta y[i] + \sum_{i:\mathcal{M}_{ii}=1} \mathbb{E}[z_j[i] \mid M_{ii}=1] \Delta y[i] && (105)\\
&= \sum_{i:\mathcal{M}_{ii}=1} \widetilde{z}_{j,\mathcal{M}}[i] \Delta y[i] + \sum_{i:\mathcal{M}_{ii}=1} \mathbb{E}[z_j[i] \mid M_{ii}=1] \Delta y[i], && (106)
\end{aligned}$$

where $\widetilde{z}_{j,\mathcal{M}}[i]$ is now a zero-mean random variable conditioned on the realization $\mathcal{M}$.

First, we bound the term $\sum_{i:\mathcal{M}_{ii}=1} \widetilde{z}_{j,\mathcal{M}}[i] \Delta y[i]$. We collect the elements corresponding to indices where $\mathcal{M}_{ii} = 1$ into the vectors $\mathbf{z}'_{j,\mathcal{M}}$ and $\boldsymbol{\Delta} y'_{\mathcal{M}}$, which are both length $\mathrm{Tr}(\mathcal{M})$ (Figure 3 shows an example of collecting elements).

$$\underbrace{\begin{bmatrix}1\\2\\3\\4\\5\end{bmatrix}}_{\widetilde{\mathbf{z}}_{j,\mathcal{M}}}, \underbrace{\begin{bmatrix}1\\0\\-1\\1\\0\end{bmatrix}}_{\boldsymbol{\Delta} y} \rightarrow \underbrace{\begin{bmatrix}1\\3\\4\end{bmatrix}}_{\mathbf{z}'_{j,\mathcal{M}}}, \underbrace{\begin{bmatrix}1\\-1\\1\end{bmatrix}}_{\boldsymbol{\Delta} y'_{\mathcal{M}}}$$

Figure 3: An example of collecting elements at indices where $\mathcal{M}_{ii} = 1$ into smaller vectors of length $\mathrm{Tr}(\mathcal{M})$. Recall that $\boldsymbol{\Delta} y[i] \neq 0$ iff $\mathcal{M}_{ii} = 0$.

We can then express

$$\begin{aligned}
&\sum_{i:\mathcal{M}_{ii}=1} \widetilde{z}_{j,\mathcal{M}}[i] \Delta y[i] && (107)\\
&\quad= (\mathbf{z}'_{j,\mathcal{M}})^T \boldsymbol{\Delta} y'_{\mathcal{M}} && (108)\\
&\quad= \frac{1}{4}\left((\mathbf{z}'_{j,\mathcal{M}} + \boldsymbol{\Delta} y'_{\mathcal{M}})^T \mathbf{I}_{\mathrm{Tr}(\mathcal{M})} (\mathbf{z}'_{j,\mathcal{M}} + \boldsymbol{\Delta} y'_{\mathcal{M}}) - (\mathbf{z}'_{j,\mathcal{M}} - \boldsymbol{\Delta} y'_{\mathcal{M}})^T \mathbf{I}_{\mathrm{Tr}(\mathcal{M})} (\mathbf{z}'_{j,\mathcal{M}} - \boldsymbol{\Delta} y'_{\mathcal{M}})\right), && (109)
\end{aligned}$$

where we added and subtracted terms in the last equality.

We prove via the subsequent propositions that conditioned on the realization $\mathcal{M}$, the entries of $\mathbf{z}'_{j,\mathcal{M}} \pm \boldsymbol{\Delta} y'_{j,\mathcal{M}}$ are i.i.d. and sub-Gaussian with bounded norm. Thus, they satisfy the requirements to apply the Hanson-Wright inequality from Rudelson and Vershynin [65] to bound the two quadratic forms in the above expression (109).

Proposition D.2. *Conditioned on the realization $\mathcal{M}$, $z'_{j,\mathcal{M}}[i']$ has sub-Gaussian norm at most* 6.

Proof. Let i be the original index from which $z'_{j,\mathcal{M}}[i']$ was sampled.

If $j > k$, then $z'_{j,\mathcal{M}}[i'] = \widetilde{z}_{j,\mathcal{M}}[i] = z_j[i]$ irrespective of the realization $\mathcal{M}$ because feature j is not used in the comparison to determine the class label and is independent to y_α and y_β (and thus independent to $\mathbf{M}$). Further, $z_j[i]$ is simply a Gaussian (and therefore sub-Gaussian with sub-Gaussian norm $\|z_j[i]\|_{\psi_2} \leq 2$. Here we use the definition of sub-Gaussian norm from (74) reproduced here for convenience:

The sub-Gaussian norm of a random variable ξ is given by,

$$\|\xi\|_{\psi_2} = \inf_{K>0} K \tag{110}$$

$$\text{s.t. } \mathbb{E}\exp\left(\xi^2/K^2\right) \leq 2. \tag{111}$$

Otherwise, if j is one of the k features that define classes, since

$$\begin{aligned} z'_{j,\mathcal{M}}[i'] &= \widetilde{z}_{j,\mathcal{M}}[i] && (112)\\ &= z_j[i] - \mathbb{E}[z_j[i] \mid M_{ii} = 1], && (113) \end{aligned}$$

the triangle inequality states that

$$\|\widetilde{z}_{j,\mathcal{M}}[i]\|_{\psi_2} \leq \|z_j[i]\|_{\psi_2} + \|\,\mathbb{E}[z_j[i] \mid M_{ii} = 1]\|_{\psi_2}. \tag{114}$$

Note that the distribution of $z_j[i]$ conditioned on realization $\mathcal{M}$ is equivalent to the distribution obtained by conditioning on the event $M_{ii} = 1$. So it is sufficient to compute these sub-Gaussian norms conditioned on the event $M_{ii} = 1$.

We will first bound $\|z_j[i]\|_{\psi_2}$. Let $\mathcal{E}_j$ be the event that $z_j[i]$ is the maximum out of the first k features, and let $\mathcal{E}_j^c$ be the complementary event.

First, without conditioning on $\mathcal{E}_j$, we know by well-known results for the standard Gaussian that

$$\mathbb{E}\exp(\mathbf{z}_j[i]^2/5) = \sqrt{\frac{5}{3}} \leq \frac{4}{3}. \tag{115}$$

Using the law of iterated expectation we can relate this to the expectation conditioned on the events $\mathcal{E}_j$ and $\mathcal{E}_j^c$, noting that $P(\mathcal{E}_j) = 1/k$:

$$\begin{aligned} \frac{4}{3} &\geq \mathbb{E}\exp(\mathbf{z}_j[i]^2/5) && (116)\\ &= P(\mathcal{E}_j)\mathbb{E}\exp(\mathbf{z}_j[i]^2/5|\mathcal{E}_j) + P(\mathcal{E}_j^c)\mathbb{E}\exp(\mathbf{z}_j[i]^2/5|\mathcal{E}_j^c) && (117)\\ &= \frac{1}{k}\mathbb{E}\exp(\mathbf{z}_j[i]^2/5|\mathcal{E}_j) + \frac{k-1}{k}\mathbb{E}\exp(\mathbf{z}_j[i]^2/5|\mathcal{E}_j^c). && (118) \end{aligned}$$

Rearranging terms, we obtain,

$$\begin{aligned} \frac{k-1}{k}\mathbb{E}\exp(\mathbf{z}_j[i]^2/5|\mathcal{E}_j^c) &\leq \frac{4}{3} - \frac{1}{k}\mathbb{E}\exp(\mathbf{z}_j[i]^2/5|\mathcal{E}_j) && (119)\\ \implies \mathbb{E}\exp(\mathbf{z}_j[i]^2/5|\mathcal{E}_j^c) &\leq \frac{k}{k-1}\left(\frac{4}{3} - \frac{1}{k}\mathbb{E}\exp(\mathbf{z}_j[i]^2/5|\mathcal{E}_j)\right) && (120)\\ &\leq \frac{k}{k-1}\cdot\frac{4}{3} && (121)\\ &\leq 2, && (122) \end{aligned}$$

where in the second to last inequality we used the non-negativity of $\mathbb{E}\exp(\mathbf{z}_j[i]^2/5|\mathcal{E}_j)$ and in the last equality we assumed $k \geq 3$. We then have

$$\begin{aligned} \mathbb{E}\exp(\mathbf{z}_j[i]^2/5|\mathcal{E}_j^c) &= \sum_{m\neq j}\mathbb{E}\exp(\mathbf{z}_j[i]^2/5|\mathcal{E}_j^c \cap \mathcal{E}_m)P(\mathcal{E}_m \mid \mathcal{E}_j^c) && (123)\\ &= \frac{1}{k-1}\sum_{m\neq j}\mathbb{E}\exp(\mathbf{z}_j[i]^2/5|\mathcal{E}_j^c \cap \mathcal{E}_m) && (124) \end{aligned}$$

where the last equality follows by symmetry. Further by symmetry, all the terms in the above summation that we are averaging are equal, so we can express it as an average of just the terms corresponding to $m = \alpha$ and $m = \beta$, as follows:

$$\begin{aligned} (124) &= \frac{1}{2}\mathbb{E}\exp(\mathbf{z}_j[i]^2/5|\mathcal{E}_j^c \cap \mathcal{E}_\alpha) + \frac{1}{2}\mathbb{E}\exp(\mathbf{z}_j[i]^2/5|\mathcal{E}_j^c \cap \mathcal{E}_\beta) && (125)\\ &= P(\mathcal{E}_\alpha \mid \mathcal{E}_j^c \cap (\mathcal{E}_\alpha \cup \mathcal{E}_\beta))\mathbb{E}\exp(\mathbf{z}_j[i]^2/5|\mathcal{E}_j^c \cap \mathcal{E}_\alpha)\\ &\quad + P(\mathcal{E}_\beta \mid \mathcal{E}_j^c \cap (\mathcal{E}_\alpha \cup \mathcal{E}_\beta))\mathbb{E}\exp(\mathbf{z}_j[i]^2/5|\mathcal{E}_j^c \cap \mathcal{E}_\beta), && (126) \end{aligned}$$

again by symmetry. Since exactly one of $\mathcal{E}_\alpha$ and $\mathcal{E}_\beta$ are true when conditioned on $\mathcal{E}_j^c \cap (\mathcal{E}_\alpha \cup \mathcal{E}_\beta)$, we can rewrite the above as our desired expectation

$$\begin{aligned}(126) &= \mathbb{E}\exp(\mathbf{z}_j[i]^2/5|\mathcal{E}_j^c \cap (\mathcal{E}_\alpha \cup \mathcal{E}_\beta)) && (127)\\ &= \mathbb{E}\exp(\mathbf{z}_j[i]^2/5|\mathcal{E}_\alpha \cup \mathcal{E}_\beta) && (128)\\ &= \mathbb{E}\exp(\mathbf{z}_j[i]^2/5|M_{ii} = 1), && (129)\end{aligned}$$

since $M_{ii} = 1$ is equivalent to the event $\mathcal{E}_\alpha \cup \mathcal{E}_\beta$. Thus, conditioned on the event $M_{ii} = 1$, $\|z_j[i]\|_{\psi_2} \le \sqrt{5}$.

Next we consider $\| \mathbb{E}[z_j[i] \mid M_{ii} = 1]\|_{\psi_2}$. By a similar argument to above, we have that $\mathbb{E}[z_j[i] \mid M_{ii} = 1] = \mathbb{E}[z_j[i] \mid \mathcal{E}_j^c]$, so we will focus on the second quantity instead. Bounds on the max of Gaussians (Lemma C.3) state that:

$$\begin{aligned} & 0 < \mathbb{E}[\mathbf{z}_j[i] \mid \mathcal{E}_j] \le \sqrt{2\log(k)} && (130)\\ \implies\quad & 0 > \mathbb{E}[\mathbf{z}_j[i] \mid \mathcal{E}_j^c] \ge -\frac{1}{k-1}\sqrt{2\log(k)} \ge -2 && (131)\\ \implies\quad & \exp\left(\frac{\mathbb{E}[\mathbf{z}_j[i] \mid \mathcal{E}_j^c]^2}{3^2}\right) < 2. && (132)\end{aligned}$$

In the second last inequality we use the fact that the function $f(k) = \left|\sqrt{2\log k}/(k-1)\right|$ is monotonically decreasing in k and assumed $k \ge 3$.

Thus, the (constant) random variable $\mathbb{E}[\mathbf{z}_j[i] \mid M_{ii} = 1]$ is sub-Gaussian with parameter 3. So, by the triangle inequality, conditioned on $M_{ii} = 1$

$$\begin{aligned}\|\widetilde{z}_{j,m}[i]\|_{\psi_2} &\le \|z_j[i]\|_{\psi_2} + \|\mathbb{E}[\widetilde{z}_{j,m}\|_{\psi_2} && (133)\\ &\le \sqrt{5} + 3 && (134)\\ &\le 6. && (135)\end{aligned}$$

This completes the proof that conditioned on the realization $\mathcal{M}$, $z'_{j,\mathcal{M}}[i']$ is sub-Gaussian with norm at most 6. □

We can now prove our target result:

Proposition D.3. *With probability at least* $(1 - 6/(ndk))$,

$$\left|\mathbf{z}_j^\top \mathbf{\Delta} y\right| \le c_6\sqrt{\frac{n}{k}} \cdot \sqrt{\log(ndk)}. \qquad (136)$$

for universal constant c_6.

Proof. Our strategy will be to bound $\mathbf{z}_j^\top \mathbf{\Delta} y = \mathbf{z}_j^\top \mathbf{M}\mathbf{\Delta} y$ for every typical realization $\mathcal{M}$ of $\mathbf{M}$ that satisfies Proposition D.1. Recall that for a given realization $\mathcal{M}$ we have,

$$\mathbf{z}_j^T \mathcal{M}\mathbf{\Delta} y = \sum_{i:\mathcal{M}_{ii}=1} \widetilde{z}_{j,\mathcal{M}}[i]\Delta y[i] + \sum_{i:\mathcal{M}_{ii}=1} \mathbb{E}[z_j[i] \mid M_{ii} = 1]\Delta y[i]. \qquad (137)$$

We will use Hanson-Wright to bound the first term, which we previously expressed in (109) as:

$$\begin{aligned}&\sum_{i:\mathcal{M}_{ii}=1} \widetilde{z}_{j,\mathcal{M}}[i]\Delta y[i] && (138)\\ &\quad= \frac{1}{4}\left((\mathbf{z}'_{j,\mathcal{M}} + \mathbf{\Delta} y'_{\mathcal{M}})^T \mathbf{I}_{\mathrm{Tr}(\mathcal{M})}(\mathbf{z}'_{j,\mathcal{M}} + \mathbf{\Delta} y'_{\mathcal{M}}) - (\mathbf{z}'_{j,\mathcal{M}} - \mathbf{\Delta} y'_{\mathcal{M}})^T \mathbf{I}_{\mathrm{Tr}(\mathcal{M})}(\mathbf{z}'_{j,\mathcal{M}} - \mathbf{\Delta} y'_{\mathcal{M}})\right). && (139)\end{aligned}$$

By Proposition D.2, the sub-Gaussian conditions for the entries of $\mathbf{z}'_{j,m}$ are satisfied. Further, $\mathbf{\Delta} y'_{\mathcal{M}}$ is bounded in $[-1, 1]$, so $\|\mathbf{\Delta} y'_{\mathcal{M}}\|_{\psi_2} \le 2$. Thus, by the triangle inequality, the sub-Gaussian norm of the entries of $\mathbf{z}'_{j,\mathcal{M}} \pm \mathbf{\Delta} y'_{\mathcal{M}}$ is bounded by $K \le 6 + 2 = 8$. Also note that conditioned on the realization $\mathcal{M}$, $\mathbf{z}'_{j,\mathcal{M}}$ is zero-mean by construction and $\boldsymbol{\Delta} y'_{\mathcal{M}}$ is zero-mean by symmetry between α and β, so we can now apply the Hanson-Wright inequality to both terms.

We choose parameter

$$t = \frac{K^2}{\sqrt{c}} \sqrt{\mathrm{Tr}(\mathcal{M})} \sqrt{\log(ndk)}. \tag{140}$$

where c is the constant from the Hanson-Wright result.

So

$$\frac{t^2}{K^4 \|\mathbf{I}_{\mathrm{Tr}(\mathcal{M})}\|_{\mathsf{F}}^2} = \frac{1}{c} \log(ndk) \tag{141}$$

$$\frac{t}{K^2 \|\mathbf{I}_{\mathrm{Tr}(\mathcal{M})}\|_{\mathsf{op}}} = \frac{1}{\sqrt{c}} \sqrt{\mathrm{Tr}(\mathcal{M})} \sqrt{\log(ndk)} > \frac{1}{c} \log(ndk). \tag{142}$$

The last inequality follows since with high probability $\mathrm{Tr}(\mathcal{M}) = \Theta(\sqrt{n/k})$, by Proposition D.1, $\sqrt{\mathrm{Tr}(\mathcal{M})}\sqrt{\log(ndk)} = \Theta(\sqrt{n \log(ndk)/k})$ grows faster than $\log(ndk)$.

Finally, note that:

$$\begin{aligned}
\mathbb{E}[(\mathbf{z}'_{j,\mathcal{M}})^T \boldsymbol{\Delta} y'_{\mathcal{M}} \mid \mathbf{M} = \mathcal{M}] &= \sum_{i:\mathcal{M}_{ii}=1} \mathbb{E}[\widetilde{z}_{j,\mathcal{M}}[i] \Delta y[i] \mid \mathbf{M} = \mathcal{M}] && (143)\\
&= \sum_{i:\mathcal{M}_{ii}=1} \mathbb{E}[\widetilde{z}_{j,\mathcal{M}}[i] \Delta y[i] \mid M_{ii} = 1] && (144)\\
&= \sum_{i:\mathcal{M}_{ii}=1} \frac{1}{2} \mathbb{E}[\widetilde{z}_{j,\mathcal{M}}[i] \mid \Delta y[i] = 1] - \frac{1}{2} \mathbb{E}[\widetilde{z}_{j,\mathcal{M}}[i] \mid \Delta y[i] = -1] && (145)\\
&= 0, && (146)
\end{aligned}$$

where the last equation follows by symmetry. Knowing which of $\mathbf{z}_\alpha[i]$ or $\mathbf{z}_\beta[i]$ was the maximum does not change the conditional expectation of $\widetilde{z}_{j,\mathcal{M}}[i]$.

So, applying Hanson-Wright, with probability at least $(1 - 4/(ndk))$ we have

$$-\frac{K^2}{2} c_5 \sqrt{\mathrm{Tr}(\mathcal{M})} \sqrt{\log(ndk)} = -\frac{t}{2} \leq \widetilde{\mathbf{z}}_{j,m}^T \Delta \mathbf{y} \leq \frac{t}{2} = \frac{K^2}{2} c_5 \sqrt{\mathrm{Tr}(\mathcal{M})} \sqrt{\log(ndk)}, \tag{147}$$

where $c_5 = \frac{1}{\sqrt{c}}$.

We next consider the second term $\sum_{i:\mathcal{M}_{ii}=1} \mathbb{E}[z_j[i] \mid M_{ii} = 1] \Delta y[i]$ from (106) conditioned on the realization $\mathcal{M}$. By an identical symmetry argument as for the previous term we have, $0 \geq \mathbb{E}[z_j[i] \mid \mathcal{E}_j^c] = \mathbb{E}[z_j[i] \mid M_{ii} = 1]$. Then as a consequence of Lemma C.3 and using the fact that $M_{ii} = 1$ implies $z_j[i]$ is not the maximum of k Gaussians we have, $\mathbb{E}[z_j[i] \mid \mathcal{E}_j^c] \geq -2\sqrt{\log(k)}/(k-1)$. So we can bound

$$\left| \sum_{i:\mathcal{M}_{ii}=1} \mathbb{E}[z_j[i] \mid M_{ii} = 1] \Delta y[i] \right| \leq \frac{2\sqrt{\log(k)}}{k-1} \left| \sum_{i:\mathcal{M}_{ii}=1} \boldsymbol{\Delta} y_i \right| \leq \frac{2\delta' \sqrt{\log(k)}}{k-1}, \tag{148}$$

with probability $1 - 2e^{-\delta'^2/(6 \cdot \mathrm{Tr}(\mathcal{M}))}$, by application of the Chernoff bound and using the fact that conditioned on $M_{ii} = 1$, $\Delta y[i]$ takes value ± 1 with probability half by symmetry among features α and β.

Next, we apply the high probability bounds above on typical realizations $\mathcal{M}$. In particular, we substitute bounds on $\mathrm{Tr}(\mathbf{M})$ from (102) from Proposition D.1 with $\delta = 1/2$ into (148), and set $\delta' = \sqrt{6(1+\delta)(n/k)\log(ndk)}$. Then $e^{-\delta'^2/(6 \cdot \mathrm{Tr}(\mathbf{M}))} \leq 1/(ndk)$ and $e^{-\frac{2n\delta^2}{3k}} < 1/(ndk)$, so using the union bound we have with probability at least $(1 - 4/(ndk) - 1/(ndk) - 1/(ndk))$,

$$
\begin{aligned}
\left|\mathbf{z}_j^T \Delta \mathbf{y}\right| &\le \left|\sum_{i:\mathcal{M}_{ii}=1} \widetilde{z}_{j,\mathcal{M}}[i]\Delta y[i]\right| + \left|\sum_{i:\mathcal{M}_{ii}=1} \mathbb{E}[z_j[i] \mid M_{ii}=1]\Delta y[i]\right| && (149)\\
&\le \frac{K^2}{2} c_5 \sqrt{1+\delta} \cdot \sqrt{\frac{2n}{k}} \cdot \sqrt{\log(ndk)} + \frac{2\sqrt{(1+\delta)(n/k)\log(ndk)}\sqrt{\log(k)}}{k-1} && (150)\\
&\le \frac{K^2}{2} c_5 \sqrt{1+\delta} \cdot \sqrt{\frac{2n}{k}} \cdot \sqrt{\log(ndk)} + \frac{2\sqrt{(1+\delta)}\sqrt{\log(k)}}{k-1} \cdot \sqrt{\frac{n}{k}} \cdot \sqrt{\log(ndk)} && (151)\\
&\le c_6 \sqrt{\frac{n}{k}} \cdot \sqrt{\log(ndk)}, && (152)
\end{aligned}
$$

for a suitable choice of c_6. □

D.1.2 Bounds on $\mathbf{z}_j^\top \mathbf{A}^{-1}(\mathbf{y}_\alpha - \mathbf{y}_\beta)$

We can now prove bounds on our target quantity. We restate the lemma that we are trying to prove below for convenience.

Lemma D.1. *Let $\boldsymbol{\Delta} y = \mathbf{y}_\alpha - \mathbf{y}_\beta$. Let α, β, and j be distinct. Then, with probability at least $(1 - 7/(ndk))$, we have,*

$$
\left|\mathbf{z}_j^\top \mathbf{A}^{-1} \boldsymbol{\Delta} y\right| \le c_7(\bar{\mu}\sqrt{\frac{n}{k}} \cdot \sqrt{\ln(ndk)} + \Delta_\mu \cdot n/\sqrt{k}), \tag{96}
$$

for some constant c_7.

Proof. We can rewrite

$$
\begin{aligned}
\mathbf{z}_j^\top \mathbf{A}^{-1} \boldsymbol{\Delta} y &= \mathbf{z}_j^\top \left(\bar{\mu}\mathbf{I}_n + \boldsymbol{\Delta} A_{inv}\right) \boldsymbol{\Delta} y && (153)\\
&= \bar{\mu}\mathbf{z}_j^\top \boldsymbol{\Delta} y + \mathbf{z}_j^\top \boldsymbol{\Delta} A_{inv} \boldsymbol{\Delta} y. && (154)
\end{aligned}
$$

Next we can bound $\left|\mathbf{z}_j^\top \boldsymbol{\Delta} A_{inv} \boldsymbol{\Delta} y\right|$ simply as

$$
\begin{aligned}
|\mathbf{z}_j^\top \boldsymbol{\Delta} A_{inv} \boldsymbol{\Delta} y| &\le \|\mathbf{z}_j\|_2 \|\boldsymbol{\Delta} A_{inv} \boldsymbol{\Delta} y\|_2 && (155)\\
&\le \|\boldsymbol{\Delta} A_{inv}\|_{op} \|\mathbf{z}_j\|_2 \|\boldsymbol{\Delta} y\|_2 && (156)\\
&\le \Delta_\mu \|\mathbf{z}_j\|_2 \|\boldsymbol{\Delta} y\|_2, && (157)
\end{aligned}
$$

where we use the fact that $\boldsymbol{\Delta} A_{inv}$ is a symmetric matrix and its 2-norm is its maximum absolute eigenvalue. We obtain the eigenvalue bounds for $\boldsymbol{\Delta} A_{inv}$ from Lemma B.1, holding with probability at least $1 - 2e^{-n}$.

So, by the triangle inequality, we have with probability at least $(1 - 6/(ndk) - 2e^{-n} - 2e^{-\frac{2n\delta^2}{3k}} - 2e^{-n\delta^2})$

$$
\left|\mathbf{z}_j^\top \mathbf{A}^{-1} \boldsymbol{\Delta} y\right| \le c_6 \bar{\mu}\sqrt{\frac{n}{k}} \cdot \sqrt{\ln(ndk)} + \Delta_\mu \cdot \sqrt{(1+\delta)n} \cdot \sqrt{(1+\delta)\frac{2n}{k}}. \tag{158}
$$

The first term follows from Proposition D.3, and the second from our bound on $\mathrm{Tr}(\mathbf{M}) = \|\boldsymbol{\Delta} y\|_2^2$ from Proposition D.1, as well as an analogous application of the chi-squared bound (Lemma C.2) on $\|\mathbf{z}_j\|_2$.

The proof follows by setting δ to any value in $(0, 1)$, choosing an appropriate constant c_7, and noting that for large enough n, $1/(ndk) \gg d_1 e^{-d_2 n/k}$ for any positive constants d_1, d_2. □

D.2 Proof of Lemma D.2

Next we use a similar technique as in Appendix D.1 to bound $\mathbf{z}_\alpha^\top \mathbf{A}^{-1} \boldsymbol{\Delta} y$. We will write $\mathbf{A}^{-1} = \bar{\mu}\mathbf{I}_n + \boldsymbol{\Delta} A_{inv}$, and split up the expression $\mathbf{z}_\alpha^\top \mathbf{A}^{-1} \boldsymbol{\Delta} y$ into components involving $\bar{\mu}\mathbf{I}_n$, and components involving $\boldsymbol{\Delta} A_{inv}$.

Proposition D.4. *Consider two arbitrary length-n zero-mean vectors $\mathbf{y}$ and $\mathbf{z}$ whose components each has sub-Gaussian norm at most K. With probability at least $1 - 4/(nk)$ we have each of*

$$\mathbf{z}^\top \mathbf{y} \leq \mathbb{E}[\mathbf{z}^\top \mathbf{y}] + 2c_8\sqrt{n} \cdot \sqrt{\ln(nk)} \tag{159}$$

$$\mathbf{z}^\top \mathbf{y} \geq \mathbb{E}[\mathbf{z}^\top \mathbf{y}] - 2c_8\sqrt{n} \cdot \sqrt{\ln(nk)}, \tag{160}$$

for some universal constant c_8.

Proof. The upper-bound follows as

$$\mathbf{z}^\top \mathbf{y} = \frac{1}{4}\left((\mathbf{z}+\mathbf{y})^\top(\mathbf{z}+\mathbf{y}) - (\mathbf{z}-\mathbf{y})^\top(\mathbf{z}-\mathbf{y})\right) \tag{161}$$

$$\leq \mathbb{E}[\mathbf{z}^\top \mathbf{y}] + \frac{K^2}{2\sqrt{c}}\sqrt{n} \cdot \sqrt{\ln nk}, \tag{162}$$

with probability at least $(1 - 4/(nk))$, where we apply the Hanson-Wright inequality to each of the quadratic terms with $t = \frac{K^2}{\sqrt{c}}\sqrt{n}\sqrt{\ln(nk)}$ and use the fact that, letting $\mathbf{M} = \mathbf{I}_n$, $\|\mathbf{M}\|_F^2 = n, \|\mathbf{M}\|_{op} = 1$. The lower-bound can be obtained analogously, and an appropriate choice of c_8 completes the proof. □

From this, we can now prove Lemma D.2, restated below for convenience:

Lemma D.2. *Let $\mathbf{\Delta}y = \mathbf{y}_\alpha - \mathbf{y}_\beta$. With probability at least $(1 - 5/(nk))$, we have each of*

$$\mathbf{z}_\alpha^\top \mathbf{A}^{-1}\mathbf{\Delta}y \leq \bar{\mu}(\mathbb{E}[\mathbf{z}_\alpha^\top \mathbf{y}_\alpha] - \mathbb{E}[\mathbf{z}_\alpha^\top \mathbf{y}_\beta]) + c_9(\bar{\mu}\sqrt{n}\sqrt{\ln(nk)} + \Delta_\mu \cdot n/\sqrt{k}) \tag{97}$$

$$\mathbf{z}_\alpha^\top \mathbf{A}^{-1}\mathbf{\Delta}y \geq \bar{\mu}(\mathbb{E}[\mathbf{z}_\alpha^\top \mathbf{y}_\alpha] - \mathbb{E}[\mathbf{z}_\alpha^\top \mathbf{y}_\beta]) - c_9(\bar{\mu}\sqrt{n}\sqrt{\ln(nk)} + \Delta_\mu \cdot n/\sqrt{k}), \tag{98}$$

for some constant c_9.

Proof. We have

$$\mathbf{z}_\alpha^\top \mathbf{A}^{-1}(\mathbf{y}_\alpha - \mathbf{y}_\beta) = \mathbf{z}_\alpha^\top \left(\bar{\mu}\mathbf{I}_n + \mathbf{\Delta}A_{inv}\right)(\mathbf{y}_\alpha - \mathbf{y}_\beta) \tag{163}$$

$$= \bar{\mu}\mathbf{z}_\alpha^\top(\mathbf{y}_\alpha - \mathbf{y}_\beta) + \mathbf{z}_\alpha^\top \mathbf{\Delta}A_{inv}(\mathbf{y}_\alpha - \mathbf{y}_\beta) \tag{164}$$

$$= \bar{\mu}\mathbf{z}_\alpha^\top(\mathbf{y}_\alpha - \mathbf{y}_\beta) + \mathbf{z}_\alpha^\top \mathbf{\Delta}A_{inv}(\mathbf{y}_\alpha^{oh} - \mathbf{y}_\beta^{oh}). \tag{165}$$

We again simply bound

$$|\mathbf{z}_\alpha^\top \mathbf{\Delta}A_{inv}\mathbf{\Delta}y| \leq \|\mathbf{z}_j\|_2\|\mathbf{\Delta}A_{inv}\mathbf{\Delta}y\|_2 \tag{166}$$

$$\leq \|\mathbf{\Delta}A_{inv}\|_{op}\|\mathbf{z}_\alpha\|_2\|\mathbf{\Delta}y\|_2 \tag{167}$$

$$\leq \Delta_\mu\|\mathbf{z}_\alpha\|_2\|\mathbf{\Delta}y\|_2 \tag{168}$$

$$\leq \Delta_\mu \cdot \sqrt{(1+\delta)n} \cdot \sqrt{(1+\delta)\frac{2n}{k}} \tag{169}$$

$$= \Delta_\mu(1+\delta)\sqrt{2}\frac{n}{\sqrt{k}}, \tag{170}$$

with probability $(1 - 2e^{-n\delta^2} - 2e^{-\frac{2n\delta^2}{3k}})$, using chi-squared bounds for $\mathbf{z}_\alpha$ (Lemma C.2) and Chernoff bounds for $\mathbf{\Delta}y$ (Proposition D.1).

With probability $(1 - 2e^{-n\delta^2} - 2e^{-\frac{2n\delta^2}{3k}})$, we get each of

$$\mathbf{z}_\alpha^T \mathbf{A}^{-1}(\mathbf{y}_\alpha - \mathbf{y}_\beta) \leq \bar{\mu}\mathbf{z}_\alpha^T(\mathbf{y}_\alpha - \mathbf{y}_\beta) + \Delta_\mu(1+\delta)\sqrt{2}\frac{n}{\sqrt{k}} \tag{171}$$

$$\mathbf{z}_\alpha^T \mathbf{A}^{-1}(\mathbf{y}_\alpha - \mathbf{y}_\beta) \geq \bar{\mu}\mathbf{z}_\alpha^T(\mathbf{y}_\alpha - \mathbf{y}_\beta) - \Delta_\mu(1+\delta)\sqrt{2}\frac{n}{\sqrt{k}}. \tag{172}$$

By applying Proposition D.4 on the relevant terms, setting δ to be an arbitrary value in $(0, 1)$, and choosing an appropriate constant c_9, we obtain with probability $(1 - 5/(nk))$ each of

$$\mathbf{z}_\alpha^T \mathbf{A}^{-1}(\mathbf{y}_\alpha - \mathbf{y}_\beta) \leq \bar{\mu}(\mathbb{E}[\mathbf{z}_\alpha^T \mathbf{y}_\alpha] - \mathbb{E}[\mathbf{z}_\alpha^T \mathbf{y}_\beta]) + c_9(\bar{\mu}\sqrt{n}\sqrt{\ln(nk)} + \Delta_\mu n/\sqrt{k}) \tag{173}$$

$$\mathbf{z}_\alpha^T \mathbf{A}^{-1}(\mathbf{y}_\alpha - \mathbf{y}_\beta) \geq \bar{\mu}(\mathbb{E}[\mathbf{z}_\alpha^T \mathbf{y}_\alpha] - \mathbb{E}[\mathbf{z}_\alpha^T \mathbf{y}_\beta]) - c_9(\bar{\mu}\sqrt{n}\sqrt{\ln(nk)} + \Delta_\mu n/\sqrt{k}). \tag{174}$$

The probability comes from the union bound $(1 - 2e^{-n\delta^2} - 2e^{-\frac{2n\delta^2}{3k}} - 4/(nk)) \geq 1 - 5/(nk)$ (for sufficiently large n). □

D.3 Proof of Corollary D.2.1

We claim the following bound:

Proposition D.5. *Bounds on* $\mathbb{E}[\mathbf{z}_\alpha^\top \mathbf{y}_\alpha]$.

$$\frac{1}{\sqrt{\pi \ln 2}} \cdot \frac{n}{k} \cdot \sqrt{\ln k} \leq \mathbb{E}[\mathbf{z}_\alpha^\top \mathbf{y}_\alpha] \leq \sqrt{2} \cdot \frac{n}{k} \cdot \sqrt{\ln k} \tag{175}$$

Proof.

$$\begin{aligned}
\mathbb{E}[\mathbf{z}_\alpha^\top \mathbf{y}_\alpha] &= \mathbb{E}[\mathbf{z}_\alpha^\top \mathbf{y}_\alpha^{oh}] - \mathbb{E}[\mathbf{z}_\alpha^\top \frac{1}{c}\mathbf{1}] && (176)\\
&= \mathbb{E}[\mathbf{z}_\alpha^\top \mathbf{y}_\alpha^{oh}] && (177)\\
&= n \left(\mathbb{E}[z_{\alpha,i} y_{\alpha,i}^{oh} | y_{\alpha,i}^{oh} = 1] P(y_{\alpha,i}^{oh} = 1) + \mathbb{E}[z_{\alpha,i} y_{\alpha,i}^{oh} | y_{\alpha,i}^{oh} = 0] P(y_{\alpha,i}^{oh} = 0)\right) && (178)\\
&= \frac{n}{k}\mathbb{E}[z_{\alpha,i} | y_{\alpha,i}^{oh} = 1]. && (179)
\end{aligned}$$

So the desired bound follows from the bounds in Lemma C.3. □

We can obtain a similar bound for when $\beta \neq \alpha$:

Proposition D.6. *Bounds on* $\mathbb{E}[\mathbf{z}_\alpha^\top \mathbf{y}_\beta]$.

$$-\sqrt{2} \cdot \frac{n}{k} \cdot \frac{1}{k-1} \cdot \sqrt{\ln k} \leq \mathbb{E}[\mathbf{z}_\alpha^\top \mathbf{y}_\beta] \leq -\frac{1}{\sqrt{\pi \ln 2}} \cdot \frac{n}{k} \cdot \frac{1}{k-1} \cdot \sqrt{\ln k} \tag{180}$$

Proof. Observe that,

$$\begin{aligned}
\mathbb{E}[\mathbf{z}_\alpha^\top \mathbf{y}_\beta] &= \mathbb{E}[\mathbf{z}_\alpha^\top \mathbf{y}_\beta^{oh}] - \mathbb{E}[\mathbf{z}_\alpha^\top \frac{1}{k}\mathbf{1}] && (181)\\
&= \mathbb{E}[\mathbf{z}_\alpha^\top \mathbf{y}_\beta^{oh}] - \frac{1}{k}\mathbb{E}[\mathbf{z}_\alpha]^\top \mathbf{1} && (182)\\
&= \mathbb{E}[\mathbf{z}_\alpha^\top \mathbf{y}_\beta^{oh}] && (183)\\
&= \sum_i \mathbb{E}[z_{\alpha,i} y_{\beta,i}^{oh}] && (184)\\
&= n \left(\mathbb{E}[z_{\alpha,i} y_{\beta,i}^{oh} | y_{\beta,i}^{oh} = 1] P(y_{\beta,i}^{oh} = 1) + \mathbb{E}[z_{\alpha,i} y_{\beta,i}^{oh} | y_{\beta,i}^{oh} = 0] P(y_{\beta,i}^{oh} = 1)\right) && (185)\\
&= \frac{n}{k}\mathbb{E}[z_{\alpha,i} | y_{\beta,i}^{oh} = 1] && (186)
\end{aligned}$$

Now, observe that

$$\begin{aligned}
\mathbb{E}[z_{\alpha,i} | y_{\alpha,i}^{oh} = 0] &= \sum_{\beta \neq \alpha} \mathbb{E}[z_{\alpha,i} | y_{\beta,i}^{oh} = 1] \Pr(y_{\beta,i} = 1 \mid y_{\alpha,i}^{oh} = 0) && (187)\\
&= \frac{1}{k-1} \sum_{\beta \neq \alpha} \mathbb{E}[z_{\alpha,i} | y_{\beta,i}^{oh} = 1] && (188)\\
&= \mathbb{E}[z_{\alpha,i} | y_{\beta,i}^{oh} = 1] && (189)
\end{aligned}$$

for a particular β, by symmetry over the possible β.

Next we bound $\mathbb{E}[z_{\alpha,i} | y_{\alpha,i}^{oh} = 0]$ as follows:

$$\begin{aligned}
& \mathbb{E}[z_{\alpha,i} | y_{\alpha,i}^{oh} = 1] P(y_{\alpha,i}^{oh} = 1) \\
+ \quad & \mathbb{E}[z_{\alpha,i} | y_{\alpha,i}^{oh} = 0] P(y_{\alpha,i}^{oh} = 0) = \mathbb{E}[z_{\alpha,i}] = 0 && (190)\\
\implies \quad & \mathbb{E}[z_{\alpha,i} | y_{\alpha,i}^{oh} = 0] \frac{k-1}{k} = -\mathbb{E}[z_{\alpha,i} | y_{\alpha,i}^{oh} = 1] \frac{1}{k} && (191)\\
\implies \quad & \mathbb{E}[z_{\alpha,i} | y_{\alpha,i}^{oh} = 0] = -\mathbb{E}[z_{\alpha,i} | y_{\alpha,i}^{oh} = 1] \frac{1}{k-1} && (192)
\end{aligned}$$

Thus, substituting in the results from Lemma C.3, and plugging back into (186), we obtain

$$-\sqrt{2} \cdot \frac{n}{k} \cdot \frac{1}{k-1} \cdot \sqrt{\ln k} \leq \mathbb{E}[\mathbf{z}_\alpha^\top \mathbf{y}_\beta] \leq -\frac{1}{\sqrt{\pi \ln 2}} \cdot \frac{n}{k} \cdot \frac{1}{k-1} \cdot \sqrt{\ln k}, \tag{193}$$

the desired result. □

We can now prove Corollary D.2.1, which we restate below for convenience:

Corollary D.2.1. *Let* $\mathbf{\Delta} y = \mathbf{y}_\alpha - \mathbf{y}_\beta$. *With probability at least* $(1 - 5/(nk))$*, we have,*

$$\mathbf{z}_\alpha^\top \mathbf{A}^{-1} \mathbf{\Delta} y \geq c_{10} \bar{\mu} \frac{n}{k} \sqrt{\ln(k)} - c_9 (\bar{\mu} \sqrt{n} \sqrt{\ln(nk)} + \Delta_\mu \cdot n / \sqrt{k}), \tag{99}$$

for some constant c_{10}.

Proof. This follows by substituting the lower bound from (175) in Proposition D.5 and the upper bound from (180) in Proposition D.6 into (98) from Lemma D.2, making an appropriate choice for c_{10}. □

E Misclassification Events: Proof of Lemmas used in Theorem 5.1

With the previous section's utility bounds that allow us to deal with multiclass training data in hand, we are in a position to establish all the lemmas that we need to analyze misclassification.

E.1 Proof of Lemma B.2: Lower bound on $\min_\beta (X_\alpha - X_\beta)$

With these bounds in hand, we can look at each misclassification event in turn. The first event to consider is if the best competing feature is unusually close to the true (maximum) feature.

Lemma B.2. ***(Lower bound on the closest feature margin as*** $k \to \infty$***):*** *For any constant* $\varepsilon > 0$*, there exists a constant* θ *such that, for sufficiently large* k *with probability at least* $(1 - \varepsilon)$*,*

$$\min_{\beta : 1 \leq \beta \neq \alpha \leq k} (x_{test}[\alpha] - x_{test}[\beta]) \geq \frac{\theta}{\sqrt{2 \ln(k)}}. \tag{45}$$

Here, α *is fixed and corresponds to the index of the true class — i.e.* α *corresponds to the index of the maximum feature among the first* k *features.*

Proof. The following result from [34] whose proof we reproduce here[17], enables us to bound the closest feature margin as:

$$\Pr(\min_\beta (x_{test}[\alpha] - x_{test}[\beta]) > \theta / \sqrt{2 \ln(k)}) \geq c_{11} e^{-\theta}, \tag{194}$$

for some universal positive constant c_{11}, for sufficiently large k. Thus, by selecting a constant θ such that $c_{11} e^{-\theta} = 1 - \varepsilon$ and choosing a sufficiently large k, we have that with probability $(1 - \varepsilon)$:

$$\min_\beta (x_{test}[\alpha] - x_{test}[\beta]) \geq \theta / \sqrt{2 \ln k}. \tag{195}$$

The proof [34] is reproduced below, with slight adaptations to match our use-case: Let β be the index of the largest competing feature to $x_{test}[\alpha]$. Then, their joint PDF becomes

$$f(x_{test}[\beta], x_{test}[\alpha]) = k(k-1) F(x_{test}[\beta])^{k-2} f(x_{test}[\beta]) f(x_{test}[\alpha]) \mathbf{1}(x_{test}[\beta] < x_{test}[\alpha]). \tag{196}$$

where F and f are the CDF and PDF of the standard Gaussian. Let

$$x = \frac{\theta}{\sqrt{2 \ln(k)}}. \tag{197}$$

Thus,

$$\Pr(x_{test}[\alpha] - x_{test}[\beta] > x) = k(k-1) J, \tag{198}$$

[17] We do this for the convenience of the reviewers since the source we are citing is a URL online. We believe that this is in the spirit of fair use.

where J is defined as

$$J = \int_{-\infty}^{\infty} \int_{x}^{\infty} F(w)^{k-2} f(w) f(v+w) \,\mathrm{d}v \,\mathrm{d}w \tag{199}$$

$$= \int_{-\infty}^{\infty} F(w)^{k-2} f(w) \left(\int_{x}^{\infty} f(v+w) \,\mathrm{d}v \right) \,\mathrm{d}w \tag{200}$$

$$= \int_{-\infty}^{\infty} F(w)^{k-2} f(w) (1 - F(x+w)) \,\mathrm{d}w. \tag{201}$$

Substituting $u = F(w)$, we have that

$$J = \int_{0}^{1} u^{k-2} (1 - F(x + F^{-1}(u))) \,\mathrm{d}u \tag{202}$$

$$= \int_{0}^{1-\ln(k)^2/k} u^{k-2} (1 - F(x + F^{-1}(u))) \,\mathrm{d}u + \int_{1-\ln(k)^2/k}^{1-1/(k\ln(k))} u^{k-2} (1 - F(x + F^{-1}(u))) \,\mathrm{d}u$$

$$+ \int_{1-1/(k\ln(k))}^{1} u^{k-2} (1 - F(x + F^{-1}(u))) \,\mathrm{d}u, \tag{203}$$

splitting $[0, 1]$ into three intervals, and integrating separately over each one. Let the three integrals be J_1, J_2, and J_3.

We have that

$$J_1 = \int_{0}^{1-\ln(k)^2/k} u^{k-2} (1 - F(x + F^{-1}(u))) \,\mathrm{d}u \tag{204}$$

$$\leq \int_{0}^{1-\ln(k)^2/k} u^{k-2} \,\mathrm{d}u \tag{205}$$

$$\leq (1 - \ln(k)^2/k)^{k-2} \tag{206}$$

$$\leq \exp\left(-\frac{k-2}{k} \ln(k)^2 \right) \tag{207}$$

$$= o\left(\frac{1}{k^2} \right). \tag{208}$$

Similarly,

$$J_3 = \int_{1-1/(k\ln(k))}^{1} u^{k-2} (1 - F(x + F^{-1}(u))) \,\mathrm{d}u \tag{209}$$

$$= \int_{1-1/(k\ln(k))}^{1} u^{k-2} (1 - F(F^{-1}(u))) \,\mathrm{d}u \tag{210}$$

$$\leq \int_{1-1/(k\ln(k))}^{1} u^{k-2} (1 - u) \,\mathrm{d}u \tag{211}$$

$$\leq \frac{1}{k\ln(k)} \int_{1-1/(k\ln(k))}^{1} u^{k-2} \,\mathrm{d}u \tag{212}$$

$$= o\left(\frac{1}{k^2} \right). \tag{213}$$

Finally, in the intermediate interval $u \in [1 - \ln(k)^2/k, 1 - 1/(k\ln(k))]$, as $k \to \infty$, we see that $u \geq (1 - \ln(k)^2/k) \to 1$, $x \to 0$, and $F^{-1}(u) \to \infty$, so for sufficiently large k,

$$\begin{aligned} 1 - F(x + F^{-1}(u)) &\simeq \frac{f(x + F^{-1}(u))}{x + F^{-1}(u)} && (214) \\ &\simeq \frac{f(x + F^{-1}(u))}{F^{-1}(u)} && (215) \\ &\simeq \frac{f(F^{-1}(u))}{F^{-1}(u)} e^{-xF^{-1}(u)} && (216) \\ &\simeq (1 - F(F^{-1}(u)))e^{-xF^{-1}(u)} && (217) \\ &\simeq (1 - u)e^{-xF^{-1}(u)}, && (218) \end{aligned}$$

applying the well-known approximation for the Gaussian CCDF $1 - F(w) \approx f(w)/w$ for large w (for example, see Eqn. 8.2.38 from Gallager [27]), and substituting in the Gaussian PDF.

Further, since $F^{-1}(u) \to \infty$, we have that

$$\begin{aligned} & & 1 - F(F^{-1}(u)) &\simeq \frac{f(F^{-1}(u))}{F^{-1}(u)} && (219) \\ &\implies & 1 - u &\simeq \frac{e^{-(F^{-1}(u))^2/2}}{F^{-1}(u)\sqrt{2\pi}} && (220) \\ & & &\simeq e^{-(F^{-1}(u))^2/(2+o(1))} && (221) \\ &\implies & F^{-1}(u) &\simeq \sqrt{-2\ln(1-u)} && (222) \\ & & &\simeq \sqrt{2\ln(k)}, && (223) \end{aligned}$$

where the last step follows from the bounds on u in the intermediate interval.

Substituting the bounds from (218) and (223) into the expression for J_2, and applying the definition of x from (197), we have,

$$\begin{aligned} J_2 &\simeq \int_{1-\ln(k)^2/k}^{1-1/(k\ln(k))} u^{k-2}(1 - F(x + F^{-1}(u)))\,\mathrm{d}u && (224) \\ &\simeq \int_{1-\ln(k)^2/k}^{1-1/(k\ln(k))} u^{k-2}(1-u)e^{-xF^{-1}(u)}\,\mathrm{d}u && (225) \\ &\simeq \int_{1-\ln(k)^2/k}^{1-1/(k\ln(k))} u^{k-2}(1-u)e^{-(\theta/\sqrt{2\ln(k)})\sqrt{2\ln(k)}}\,\mathrm{d}u && (226) \\ &\simeq \int_{1-\ln(k)^2/k}^{1-1/(k\ln(k))} u^{k-2}(1-u)e^{-\theta}\,\mathrm{d}u && (227) \\ &\simeq e^{-\theta}\left[\frac{u^{k-1}}{k-1} - \frac{u^{k-2}}{k-2}\right]_{1-\ln(k)^2/k}^{1-1/(k\ln(k))} && (228) \\ &\simeq \frac{e^{-\theta}}{(k-1)(k-2)}. && (229) \end{aligned}$$

Combining the terms from (208), (213), and (229), and substituting back into (198), we see that

$$\begin{aligned} \Pr\left(x_{test}[\alpha] - x_{test}[\beta] > x\right) &= k(k-1)J && (230) \\ &= k(k-1)(J_1 + J_2 + J_3) && (231) \\ &\simeq k(k-1)\left(\frac{e^{-\theta}}{(k-1)(k-2)} + o\left(\frac{1}{k^2}\right)\right) && (232) \\ &\simeq e^{-\theta}. && (233) \end{aligned}$$

Expressing this as a non-asymptotic lower-bound on the probability, holding for sufficiently large k, yields the cited result in (194).

□

Lemma B.3. ***(Lower bound on the closest feature margin when k is constant)****: If $k = c_k$ for some fixed constant c_k, for any constant $\varepsilon > 0$, there exists a constant $\varepsilon' > 0$ such that*

$$\Pr\left(\min_{\beta,\gamma:1\leq\beta\neq\gamma\leq c_k} |x_{test}[\beta] - x_{test}[\gamma]| \geq \varepsilon'\right) \geq 1-\varepsilon. \tag{46}$$

Thus, with probability at least $(1-\varepsilon)$,

$$\min_{\beta:1\leq\beta\neq\alpha\leq k} (x_{test}[\alpha] - x_{test}[\beta]) \geq \varepsilon'. \tag{47}$$

Here, α is fixed and corresponds to the index of the true class — i.e. α corresponds to the index of the maximum feature among the first k features.

Proof. Observe that,

$$\min_{1\leq\beta\neq\alpha\leq c_k} (x_{test}[\alpha] - x_{test}[\beta]) \geq \min_{1\leq\beta\neq\gamma\leq c_k} |x_{test}[\beta] - x_{test}[\gamma]|. \tag{234}$$

In other words, rather than bounding the margin between the largest and second-largest features, we will lower-bound the absolute difference between any pair of features.

Consider a particular (β,γ) tuple. Observe that $x_{test}[\beta] - x_{test}[\gamma] \sim N(0,2)$, since each feature is drawn independently from a standard Gaussian. For any $\epsilon' > 0$, we can upper-bound

$$\Pr\left(|x_{test}[\beta] - x_{test}[\gamma]| \leq \varepsilon'\right) \leq \frac{\varepsilon'}{\sqrt{\pi}} \tag{235}$$

by taking the product of the maximum value of the Gaussian pdf and the width, $2\epsilon'$, of the region we are interested in. Taking the union bound across all (β,γ) tuples, we find that

$$\Pr\left(\min_{1\leq\beta\neq\gamma\leq c_k} |x_{test}[\beta] - x_{test}[\gamma]| \leq \varepsilon'\right) \leq \frac{c_k^2\varepsilon'}{\sqrt{\pi}}. \tag{236}$$

So for any given $\varepsilon > 0$, we can choose $\varepsilon' = \varepsilon\sqrt{\pi}/c_k^2$, and have that

$$\Pr\left(\min_{1\leq\beta\neq\gamma\leq c_k} |x_{test}[\beta] - x_{test}[\gamma]| \geq \varepsilon'\right) \geq 1-\varepsilon. \tag{237}$$

□

E.2 Lower bound on $\frac{\lambda\widehat{h}_{\alpha,\beta}[\alpha]}{\max_\beta \mathsf{CN}_{\alpha,\beta}}$

Next, we will find a lower bound for survival-contamination ratio within the regime with low survival variance.

Lemma B.4. ***(Lower bound on relative survival of true feature)****: For any fixed $\beta \in [k]$, $\beta \neq \alpha$, with $\lambda_\alpha = \lambda_\beta = \lambda$ we have with probability at least $(1 - 5/(nk))$,*

$$\lambda\widehat{h}_{\alpha,\beta}[\alpha] \geq \lambda\left(c_{10}\bar{\mu}\frac{n}{k}\sqrt{\ln(k)} - c_9(\bar{\mu}\sqrt{n}\sqrt{\ln(nk)} + \Delta_\mu \cdot n/\sqrt{k})\right), \tag{48}$$

for universal positive constants c_9 and c_{10}.

Proof. Using Corollary D.2.1, we lower bound $\widehat{h}_{\alpha,\beta}[\alpha]$ with probability at least $(1 - 5/(nk))$ as

$$\begin{aligned}
\widehat{h}_{\alpha,\beta}[\alpha] &= \lambda_\alpha^{-1/2}(\hat{f}_\alpha[\alpha] - \hat{f}_\beta[\alpha]) && (238)\\
&= \mathbf{z}_\alpha^\top \mathbf{A}^{-1}\mathbf{y}_\alpha - \mathbf{z}_\alpha^\top \mathbf{A}^{-1}\mathbf{y}_\beta && (239)\\
&\geq c_{10}\bar{\mu}\frac{n}{k}\sqrt{\ln(k)} - c_9(\bar{\mu}\sqrt{n}\sqrt{\ln(nk)} + \Delta_\mu \cdot n/\sqrt{k}). && (240)
\end{aligned}$$

Multiplying through by λ gives the desired result. □

From the above result, under the scalings of our bi-level model we obtain:

Corollary B.4.1. *Under the bi-level ensemble model 4.2, for any fixed $\beta \in [k]$, $\beta \neq \alpha$, $\lambda_\alpha = \lambda_\beta = \lambda$ if $t < 1/2$, $t < 2(q+r-1)$ and $1 < q+r < (p+1)/2$, with probability at least $(1 - 5/(nk))$,*

$$\lambda \widehat{h}_{\alpha,\beta}[\alpha] \geq c_{12} n^{1-q-r-t} \sqrt{\ln(k)}, \tag{49}$$

for universal positive constant c_{12}.

Proof. Substituting our asymptotic scalings into the results from Lemma B.4 and using the decay rate of $\bar{\mu} \asymp n^{-p}$ from Corollary C.1.1 (which we can do since $1 < q+r < (p+1)/2$), we find that

$$\begin{aligned}
\lambda \widehat{h}_{\alpha,\beta}[\alpha] &\geq n^{p-q-r} \left(c_{10} \bar{\mu} \frac{n}{k} \sqrt{\ln(k)} - c_9 (\bar{\mu} \sqrt{n} \sqrt{\ln(nk)} + \Delta_\mu \cdot n/\sqrt{k}) \right) && (241)\\
&= c_{10} n^{1-q-r-t} \sqrt{\ln(k)} - c_9 n^{1/2-q-r} \sqrt{\ln(nk)} - c_9 n^{2-2q-2r-t/2} && (242)\\
&\geq c_{12} n^{\max(1-q-r-t, 2-2q-2r-t/2)} \sqrt{\ln(k)} && (243)\\
&= c_{12} n^{1-q-r-t+\max(0,1-q-r+t/2)} \sqrt{\ln(k)} && (244)\\
&\geq c_{12} n^{1-q-r-t} \sqrt{\ln(k)}, && (245)
\end{aligned}$$

for an appropriately chosen universal constant c_{12} and sufficiently large n. □

Next we upper bound $\max_\beta \mathsf{CN}_{\alpha,\beta}$.

Lemma B.5. ***(Upper bound on contamination)**: For any fixed $\beta \in [k]$, $\beta \neq \alpha$, with probability at least $(1 - 7/(nk))$,*

$$\mathsf{CN}_{\alpha,\beta} \leq c_7 (\bar{\mu} \sqrt{\frac{n}{k}} \cdot \sqrt{\ln(ndk)} + \Delta_\mu \cdot n/\sqrt{k}) \sqrt{\sum \lambda_j^2}, \tag{50}$$

for universal positive constant c_7.

Proof. For each β we have,

$$\mathsf{CN}_{\alpha,\beta} = \sqrt{\left(\sum_{j \notin \{\alpha,\beta\}} \lambda_j^2 (\widehat{h}_{\beta,\alpha}[j])^2 \right)} \tag{246}$$

For $j \notin \{\alpha, \beta\}$, by Lemma D.1,

$$\begin{aligned}
\left| \widehat{h}_{\beta,\alpha}[j] \right| &= \left| \widehat{h}_{\alpha,\beta}[j] \right| && (247)\\
&= \left| \widehat{f}_j - \widehat{g}_j \right| && (248)\\
&= \left| \mathbf{z}_j^\top \mathbf{A}^{-1} \mathbf{y}_\alpha - \mathbf{z}_j^\top \mathbf{A}^{-1} \mathbf{y}_\beta \right| && (249)\\
&= \left| \mathbf{z}_j^\top \mathbf{A}^{-1} (\mathbf{y}_\alpha - \mathbf{y}_\beta) \right| && (250)\\
&\leq c_7 (\bar{\mu} \sqrt{\frac{n}{k}} \cdot \sqrt{\ln(ndk)} + \Delta_\mu \cdot n/\sqrt{k}), && (251)
\end{aligned}$$

with probability $1 - 7/(ndk)$.

So taking the union bound over all $d-2$ terms in the expression for the contamination, we can upper-bound it as

$$\mathsf{CN}_{\alpha,\beta} \leq c_7 (\bar{\mu} \sqrt{\frac{n}{k}} \cdot \sqrt{\ln(ndk)} + \Delta_\mu \cdot n/\sqrt{k}) \sqrt{\sum \lambda_j^2}, \tag{252}$$

with probability $(1 - 7/(nk))$, the desired result. □

Corollary B.5.1. *Under the bi-level model 4.2, in the regime* $1 < q+r < (p+1)/2$*, with probability at least* $(1 - 7/(nk))$*,*

$$\mathsf{CN}_{\alpha,\beta} \leq c_{13} n^{(1-t-p)/2+\max(0,3/2-q-r)+\max(0,p/2-q-r/2)} \sqrt{\ln(ndk)}, \tag{51}$$

for universal positive constant c_{13}*.*

Proof. Since $1 < q+r < (p+1)/2$, we can apply Corollary C.1.1 to the result from Lemma B.5 and substitute in the known scalings of various terms, to obtain

$$\mathsf{CN}_{\alpha,\beta} \leq c_7(n^{1/2-t/2-p}\sqrt{\ln(ndk)} + c_4 n^{2-p-q-r-t/2})(n^{p-q-r/2} + n^{p/2}) \tag{253}$$

$$\leq c_{13} n^{(1-t-p)/2+\max(0,3/2-q-r)+\max(0,p/2-q-r/2)} \sqrt{\ln(ndk)}, \tag{254}$$

for an appropriately chosen universal positive constant c_{13}.

□

E.3 Proof of Lemma B.6: Bounds on Survival Variance

Finally, we look at the error event where a competing feature has unusually high survival relative to the true feature, so it is incorrectly selected.

Lemma B.6. ***(Upper bound on survival variance)****: For any fixed competing feature* $\beta \in [k]$*,* $\beta \neq \alpha$ *with* $\lambda_\alpha = \lambda_\beta$*, we have with probability at least* $(1 - 15/(nk))$*,*

$$\frac{\widehat{h}_{\alpha,\beta}[\alpha] - \widehat{h}_{\beta,\alpha}[\beta]}{\widehat{h}_{\alpha,\beta}[\alpha]} \leq \frac{2c_9(\bar{\mu}\sqrt{n}\sqrt{\ln(nk)} + \Delta_\mu \cdot n/\sqrt{k})}{c_{10}\bar{\mu}\frac{n}{k}\sqrt{\ln(k)} - c_9(\bar{\mu}\sqrt{n}\sqrt{\ln(nk)} + \Delta_\mu \cdot n/\sqrt{k})}, \tag{52}$$

for universal positive constants c_9 *and* c_{10}*.*

Proof. We first consider the numerator of the LHS of (52). By Lemma D.2, with probability at least $(1 - 5/(nk))$,

$$\widehat{h}_{\alpha,\beta}[\alpha] = \lambda_\alpha^{-1/2}(\hat{f}_\alpha[\alpha] - \hat{f}_\beta[\alpha]) \tag{255}$$

$$= \mathbf{z}_\alpha^\top \mathbf{A}^{-1}\mathbf{y}_\alpha - \mathbf{z}_\alpha^\top \mathbf{A}^{-1}\mathbf{y}_\beta \tag{256}$$

$$\leq \bar{\mu}(\mathbb{E}[\mathbf{z}_\alpha^\top \mathbf{y}_\alpha] - \mathbb{E}[\mathbf{z}_\alpha^\top \mathbf{y}_\beta]) + c_9(\bar{\mu}\sqrt{n}\sqrt{\ln(nk)} + \Delta_\mu \cdot n/\sqrt{k}). \tag{257}$$

Similarly, with probability at least $(1 - 5/(nk))$,

$$\widehat{h}_{\beta,\alpha}[\beta] = \lambda_\beta^{-1/2}(\hat{f}_\beta[\beta] - \hat{f}_\alpha[\beta]) \tag{258}$$

$$= \mathbf{z}_\beta^\top \mathbf{A}^{-1}\mathbf{y}_\beta - \mathbf{z}_\beta^\top \mathbf{A}^{-1}\mathbf{y}_\alpha \tag{259}$$

$$\geq \bar{\mu}(\mathbb{E}[\mathbf{z}_\beta^\top \mathbf{y}_\beta] - \mathbb{E}[\mathbf{z}_\beta^\top \mathbf{y}_\alpha]) - c_9(\bar{\mu}\sqrt{n}\sqrt{\ln(nk)} + \Delta_\mu \cdot n/\sqrt{k}). \tag{260}$$

By symmetry,

$$\mathbb{E}[\mathbf{z}_\beta^\top \mathbf{y}_\beta] = \mathbb{E}[\mathbf{z}_\alpha^\top \mathbf{y}_\alpha] \tag{261}$$

$$\mathbb{E}[\mathbf{z}_\beta^\top \mathbf{y}_\alpha] = \mathbb{E}[\mathbf{z}_\alpha^\top \mathbf{y}_\beta]. \tag{262}$$

Thus with probability at least $(1 - 10/(nk))$,

$$\widehat{h}_{\alpha,\beta}[\alpha] - \widehat{h}_{\beta,\alpha}[\beta] \leq 2c_9(\bar{\mu}\sqrt{n}\sqrt{\ln(nk)} + \Delta_\mu \cdot n/\sqrt{k}). \tag{263}$$

Using Corollary D.2.1 to lower-bound the denominator of the LHS of (52), we obtain with probability at least $(1 - 15/(nk))$

$$\frac{\widehat{h}_{\alpha,\beta}[\alpha] - \widehat{h}_{\beta,\alpha}[\beta]}{\widehat{h}_{\alpha,\beta}[\alpha]} \leq \frac{2c_9(\bar{\mu}\sqrt{n}\sqrt{\ln(nk)} + \Delta_\mu \cdot n/\sqrt{k})}{c_{10}\bar{\mu}\frac{n}{k}\sqrt{\ln(k)} - c_9(\bar{\mu}\sqrt{n}\sqrt{\ln(nk)} + \Delta_\mu \cdot n/\sqrt{k})}. \tag{264}$$

□

We can apply Corollary C.1.1 to simplify our results from Lemma B.6 in the asymptotic regime for the bi-level model.

Corollary B.6.1. *Under the bi-level ensemble model 4.2, for any fixed $\beta \in [k]$, $\beta \neq \alpha$, if $t < 1/2$, $t < 2(q + r - 1)$, and $1 < q + r < (p + 1)/2$, with probability at least $(1 - 15/(nk))$,*

$$\frac{\widehat{h}_{\alpha,\beta}[\alpha] - \widehat{h}_{\beta,\alpha}[\beta]}{\widehat{h}_{\alpha,\beta}[\alpha]} < n^{-u}, \tag{53}$$

for large enough n for some fixed $u > 0$.

Proof. Substituting, using Corollary C.1.1, in the regime where $1 < q + r < (p + 1)/2$ and $t < 1/2$, we find that

$$\begin{aligned}
\frac{\widehat{h}_{\alpha,\beta}[\alpha] - \widehat{h}_{\beta,\alpha}[\beta]}{\widehat{h}_{\alpha,\beta}[\alpha]} &\leq \frac{2c_9(n^{1/2-p}\sqrt{\ln(nk)} + c_4 n^{2-p-q-r-t/2})}{c_{10}n^{1-p-t}\sqrt{\ln(k)} - c_9(n^{1/2-p}\sqrt{\ln(nk)} + c_4 n^{2-p-q-r-t/2})} && (265)\\
&\leq \frac{2c_9}{c_{10}} \cdot \frac{n^{1/2}\sqrt{\ln(nk)} + c_4 n^{2-q-r-t/2}}{n^{1-t} - (c_9/c_{10})n^{1/2}\sqrt{\ln(n)} + (c_9 \cdot c_4/c_{10})n^{2-q-r-t/2}} && (266)\\
&\leq c_{14}\frac{n^{1/2}\sqrt{\ln(nk)} + n^{2-q-r-t/2}}{n^{1-t}} && (267)\\
&\leq c_{14}n^{\max(t-1/2, t/2+1-q-r)}\sqrt{\ln(nk)}, && (268)
\end{aligned}$$

for sufficiently large n and an appropriate choice of positive constant c_{14}. Thus, if $\max(t - 1/2, t/2 + 1 - q - r) < 0$, our quantity of interest tends to zero at a polynomial rate as $n \to \infty$, completing the proof. □

F Conjectured Looseness of Bound

In (155) in the proof of Lemma D.1, we upper bound $\mathbf{z}_j^\top \mathbf{\Delta} A_{inv} \mathbf{\Delta} y$ using the Cauchy-Schwarz inequality as

$$\begin{aligned}
|\mathbf{z}_j^\top \mathbf{\Delta} A_{inv} \mathbf{\Delta} y| &\leq \|\mathbf{z}_j\|_2 \|\mathbf{\Delta} A_{inv} \mathbf{\Delta} y\|_2 && (269)\\
&\leq \|\mathbf{\Delta} A_{inv}\|_{op} \|\mathbf{z}_j\|_2 \|\mathbf{\Delta} y\|_2 && (270)\\
&\leq \Delta_\mu \|\mathbf{z}_j\|_2 \|\mathbf{\Delta} y\|_2. && (271)
\end{aligned}$$

This results in a high-probability bound of the order $\Delta_\mu n/\sqrt{k}$. Essentially this bound fears that $\mathbf{\Delta} A_{inv}$ can, in worst case, align $\mathbf{z}_j$ and $\mathbf{\Delta} y$ to be in the same direction. However, since there is only a weak dependence between $\mathbf{\Delta} A_{inv}$ and $\mathbf{z}_j$ and $\mathbf{\Delta} y$ this bound is likely overly cautious. We conjecture that this bound is loose by a factor $\sqrt{n}$. Why do we conjecture this? If we ignored the dependency of $\mathbf{\Delta} A_{inv}$ on $\mathbf{z}_j$ and $\mathbf{\Delta} y$ and blindly applied the Hanson-Wright inequality (with the $\mathbf{M}$ matrix introduced as in Appendix D.1.1 to leverage the fact that $\mathbf{\Delta} y$ is mostly zeros) then we would obtain a high-probability upper bound of the form $\Delta_\mu \sqrt{n/k}$ (ignoring the logarithmic factors).

Assuming this tighter conjectured bound holds and similarly assuming an analogously tighter bound for $\left|\mathbf{z}_\alpha^\top \mathbf{\Delta} A_{inv} \mathbf{\Delta} y\right|$ in Appendix D.2 and following through with the rest of our analysis, we obtain the conjectured sufficient conditions for good generalization as in Equation (23) from Conjecture 6.1 for the regime $q + r > 1$.

It turns out that whenever the survival/contamination ratio grows at a polynomial rate n^v for $v > 0$ then the survival variation term also shrinks at a polynomial rate n^{-u} for $u > 0$. Thus ensuring the survival/contamination ratio is large enough (i.e. the number of classes is not too large relative to the level of favoring of potentially true features) is key to obtaining good generalization.

Although we focus on the regime $q + r > 1$ in our work, our proof technique is also applicable to the regime $q + r < 1$, i.e where regression works and by grinding through the math for this setting we should be able to get sufficient conditions for good generalization here as well. The survival in the multi-class setting in the regime $q + r < 1$ will scale roughly as $1/k$ due to the fewer positive training examples per class instead of behaving like the constant $\sqrt{2/\pi}$ as was the case for binary classification (Lemma 32, [56]). Moreover, Lemma 34 from Muthukumar et al. [56] shows that for the binary classification setting the contamination scales as $n^{-\min(p-1,1-r)/2}$ when $q+r \leq 1$. In the multiclass setting the contamination will be lower by a factor of $\sqrt{k}$ and substituting this in our error analysis we obtain Conjecture 6.1 for the regime $q+r<1$.

Finally, we believe that we can adapt our analysis from the Proof of Theorem 5.1 in Appendix B to write a set of sufficient conditions for poor generalization. The primary condition for this would be for the relevant survival/contamination ratio to go to zero. We conjecture that computing conditions on p, q, r, t under which this occurs results in the converse result in the form of sufficient conditions for poor generalization present in Conjecture 6.1. Intuitively, if the survival/contamination ratio goes to zero, then the contamination can with significant probability flip the sign of a comparison involving the score that should be winning — this parallels the way that the converse is proved in Muthukumar et al. [56] for binary classification.

G Scaling parameters with the number of positive training examples per class

From our results in Figure 2 we observed that as the number of classes k increases (i.e. larger values of t), the region where multiclass classification generalizes well shrinks. A justification for this is when the number of classes k increases while the number of training points n stays constant, we have fewer positive training examples from each class, and this makes the task harder.

To see if the reduced number of positive training examples is indeed the dominant effect, we can explore what happens if we increase the number of total training points to compensate for this effect? Instead of scaling all parameters with the total number of training points, what happens if we scale them with the number of positive training examples per class?

Let $N = n^b$ be the new number of training points for some $b > 1$, while rest of the parameters in the bi-level model scale as before. We have,

$$N = n^b \tag{272}$$

$$d = n^p = N^{p/b} \tag{273}$$

$$s = n^r = N^{r/b} \tag{274}$$

$$a = n^{-q} = N^{-q/b} \tag{275}$$

$$k = c_k n^{-t} = c_k N^{-t/b}. \tag{276}$$

We can interpret this as our standard setup, albeit parameterized by N, rather than n. To keep the model well-defined we require the following:

- $b < p$, to ensure we are still overparameterized;
- $r < b$, to ensure the number of favored features does not exceed the total number of training points;
- $q < p - r$ to ensure we are actually favoring the first s features.

For this setup, Theorem 5.1 states that the probability of misclassification tends to zero if

$$\frac{t}{b} < \min\left(\frac{r}{b}, 1-\frac{r}{b}, \frac{p}{b}+1-2\left(\frac{q}{r}+\frac{r}{b}\right), \frac{p}{b}-2, \frac{2q}{b}+\frac{r}{b}-2\right) \tag{277}$$

$$\frac{q}{b}+\frac{r}{b} > 1. \tag{278}$$

Rearranging, we obtain the condition

$$t < \min(r, b-r, p+b-2(q+r), p-2b, 2q+r-2b) \tag{279}$$

$$q+r > b. \tag{280}$$

To hold the number of training samples per class fixed we can set $b = t + 1$, so the ratio N/k becomes constant. Doing so, we obtain the following sufficient conditions for good generalization:

$$t < \min\left(r, \frac{p-2}{3}, \frac{2q+r-2}{3}\right) \tag{281}$$

$$0 < 1 - r \tag{282}$$

$$0 < p + 1 - 2(q + r) \tag{283}$$

$$t < q + r - 1. \tag{284}$$

Additionally for the model to be well defined we require $t < p - 1$. (The other conditions $r < t + 1$ and $q < p - r$ for model to be well defined are automatically satisfied if the above conditions for good generalization are satisfied).

If we assume Conjecture 6.1 then a set of sufficient conditions for good generalization is:

$$0 \leq p + 1 - 2(q + r) \tag{285}$$

$$r < 1 \tag{286}$$

$$t < r \tag{287}$$

$$t < p - 1. \tag{288}$$

The first two conditions must be satisfied for binary classification problem to generalize well and thus for multi-class classification to succeed in this setting we need to ensure binary classification succeeds. The condition $t < r$ arises because if we don't favor the features used in the comparison while assigning class labels then we have no hope of succeeding in overparameterized settings. The condition $t < p - 1$ ensures that the problem is overparameterized. If any of these conditions is not met then the probability of classification error will tend to 1.

Figure 4 visualizes the conjectured regimes for this alternative setup where the number of positive training examples per class is held fixed as we vary the number of classes for fixed values of p and q. In the white region, our model is not well defined. Note that in subfigure (a), the limiting factor to the model being well defined is the inequality $r < 1 + t$ (we must have more training examples than favored features) while in subfigure (b), the limiting factor for the model being well defined in the right-hand boundary is the inequality $r < p - q$ (we must put a larger weight on the features we favor as compared to those that we do not favor). In subfigure(b) we see that the top boundary for the model being well defined is the inequality $t < p - 1$ which is necessary for the problem to be overparameterized and support the existence of interpolating solutions. Further, the right-hand bound for good generalization in subfigure (a) corresponds to the inequality $r < 1$ while in subfigure (b) it corresponds to $p + 1 > 2(q + r)$. The left-hand boundary for good generalization in both figures is the inequality $t < r$, which reflects the fact that for MNI-based classification to succeed, all the features defining the classes must be favored.

It is interesting to note that when we add more training points so as to increase the number of positive examples, we are effectively decreasing the level of overparameterization in the problem. We know from Nakkiran [58] that adding training data in a way that reduces overparameterization can sometimes make performance worse instead of better. However, in the deeply overparameterized setting of the bi-level models explored here, this effect is counteracted by the survival benefits of having more positive examples — in effect, reducing the overall level of overparameterization reduces the shrinkage induced by the regularizing effect of overparameterization. This reduction in shrinkage compensates for the $\frac{1}{k}$ hit to survival induced by the larger number of classes.

H Additional related work

H.1 Comparisons to Wang et al. [77]

Recent work [77] provides an analysis of the generalization error of the minimum-norm interpolation of one-hot labels for multiclass classification with Gaussian features. Using the bi-level model, the authors present parameter regimes where multiclass classification error goes to zero asymptotically. While our work has many similarities in terms of model and problem setting, there are some key differences.

The first key difference is in how the training data is generated. In this paper, we assume the true label of a point is generated based on which of the first k dimensions is the largest, while Wang et al.

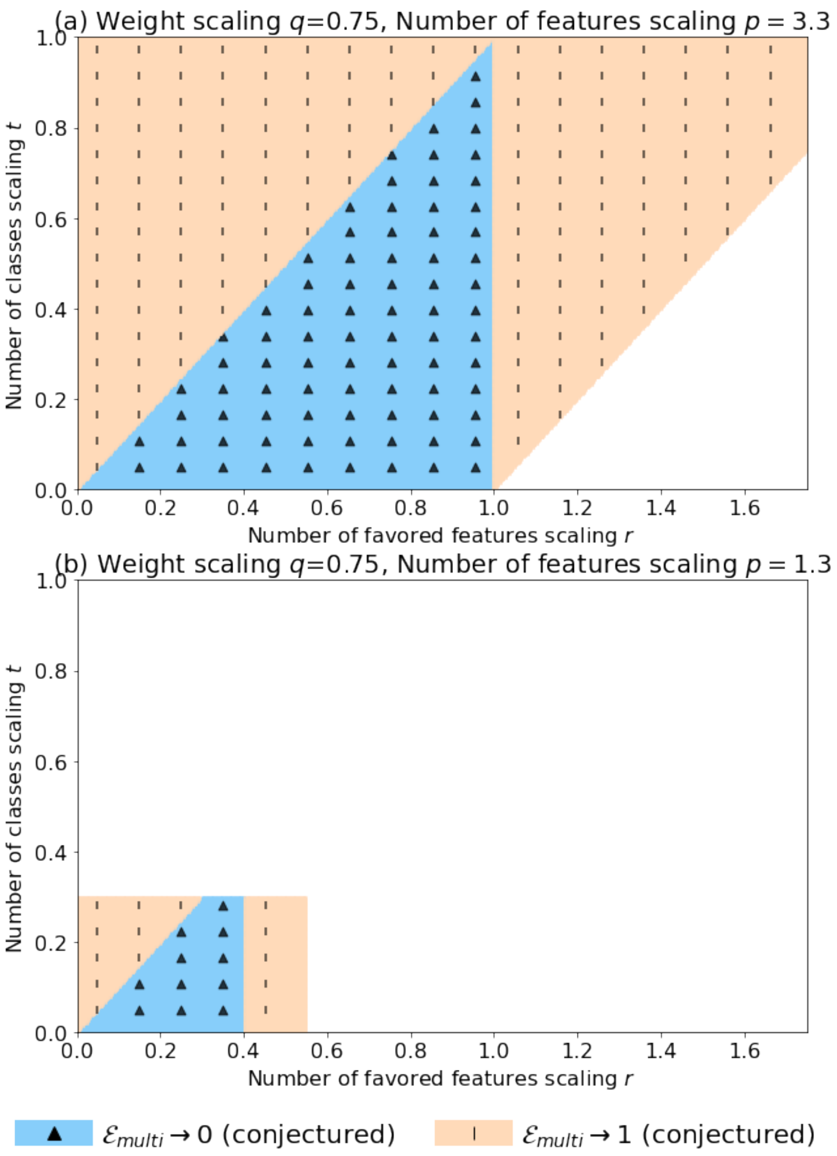


Figure 4: Visualization of the conjectured bi-level classification regimes when we scale everything with the number of positive training examples per class, instead of with the total number of training points.

[77] consider a Gaussian mixture model and a multinomial logistic model where the true labels have some randomness even conditioned on the first k dimensions. Like us, however, they also consider the case of orthogonal classes.

Second, we consider the asymptotic case where the number of classes, k scales with the number of training points as $k = cn^t$ for some positive integer c and non-negative real t. The work in Wang et al. [77] considers only the finite classes setting i.e. $t = 0$ in our model. The error analysis technique employed by us here in the form of a typicality-style argument featuring the feature margin (difference between the largest and second largest feature) is much tighter than the method employed in Wang et al. [77] and allows us to compute regimes where multiclass classification succeeds even when $t > 0$. A straight substitution into the analysis from Wang et al. [77] does not work since that analysis is too loose for this setting. Furthermore, in our expressions for survival and contamination

(Lemmas B.4 and B.5) we compute an exact dependence on k.[18] The expressions from Wang et al. [77] don't compute this exact dependence because it is not required for their purposes. By using our novel analysis technique we are able to elucidate the challenges posed by fewer positive training examples per class in the multiclass setting and provide sufficient conditions for generalization when number of classes scales with the number of training points.

An equivalence between the solution obtained by minimum-ℓ_2-norm interpolation on the adjusted zero-mean one-hot encoded labels that we perform in our approach (Equation (9)) and the solution obtained by other training methods has been established in [77]. In particular the minimum-norm interpolating solution is typically identical to the solution obtained via one-vs-all SVM and multiclass SVM (and thus gradient descent on cross-entropy loss due to its implicit bias [36, 68], under sufficient overparameterization. From Wang et al. [77], the sufficient conditions for the equivalence of solutions are,

$$\frac{\sum_{j=1}^{n} \lambda_j}{\lambda_1} > C_1 k^2 n \ln(kn), \tag{289}$$

$$\frac{(\sum_{j=1}^{n} \lambda_j)^2}{\sum_{j=1}^{n} \lambda_j^2} > C_2(\ln(kn) + n), \tag{290}$$

where C_1, C_2 are positive constants. Under our bi-level model (Definition 4.2) these conditions translate to:

$$q + r > 2t + 1, \tag{291}$$

$$2p - \max(2p - 2q - r, p) > 1. \tag{292}$$

This can be succinctly expressed as,

$$0 < t < \frac{q + r - 1}{2}. \tag{293}$$

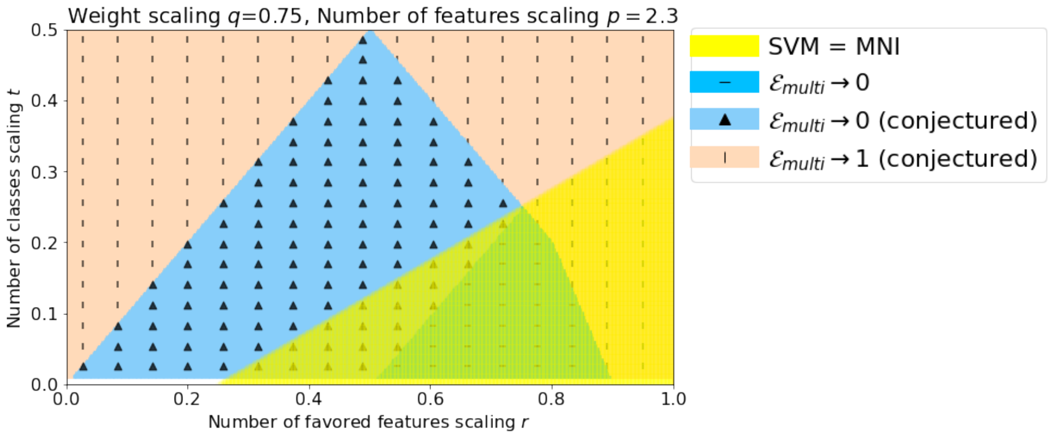


Figure 5: Visualization of regime where SVM solution is identical to MNI solution.

In Figure 5 we plot our provable as well as conjectured regimes alongside the regime where minimum-norm interpolation leads to same solution as other training methods. Notice the overlap. Thus our analysis is not limited only to the minimum-norm interpolator but holds for other training methods when the problem is sufficiently overparameterized. In this sense, the results in Wang et al. [77] and the present paper should be read together to tell a more full story of overparameterized multiclass classification.

[18] In particular, our analysis here brings out the fact that multiclass training data becomes less informative per training sample as the number of classes increases. This results in a $\frac{1}{k}$ scaling term in survival and a $\frac{1}{\sqrt{k}}$ scaling in contamination. It is this effect that makes it possible in some regimes for the contamination from other favored features to dominate — whereas in the case of binary classification, it is always the contamination from unfavored features that dominates.

H.2 Comparisons to Muthukumar et al. [56]

The work in Muthukumar et al. [56] provides an analysis of the binary classification and regression problem with Gaussian features in the overparameterized regime and shows that binary classification is easier than regression by proving the existence of a regime in a bi-level model where binary classification generalizes well but regression does not. In this work we use a similar bi-level model and the signal-processing inspired concepts of survival and contamination in our proofs but the nature of the training data in the multiclass classification problem is the key challenge and complicates our analysis considerably. Since the true class labels are generated by comparing k features, we no longer have independence of the class label y with any of these features. This is relevant when we compute bounds on the the term $z_\alpha^\top \mathbf{A}^{-1}(y_\alpha - y_\beta)$ an integral part of our survival quantity (Equations (97),(98) from Lemma D.2), since the Hanson-Wright inequality is no longer applicable directly as was the case for the binary classification problem in prior work (Appendix D.3.1 of Muthukumar et al. [56]). Working through these challenges, we prove that the multiclass problem is fundamentally different from (and harder than) than the binary problem due to the effect of fewer informative samples (positive training examples) per class. In particular we show via dominant terms from Lemmas B.6 and B.5 that, as we increase the number of classes k, survival shrinks as $1/k$ while contamination shrinks only as $1/\sqrt{k}$. Thus the survival/contamination ratio which plays a key role in the expression for classification error decreases as $1/\sqrt{k}$ in the multiclass setting as we increase k. Thus, for good generalization we need to ensure number of classes is not too large in addition to having sufficient favoring of true features.

I Experimental results

Theorem 5.1 is proved rigorously and so we know that the asymptotic result is true. Underlying the result is the analysis of survival (how strongly is the true feature underlying this class represented in the learned score) and contamination (what is the standard deviation of the contamination in predictions that comes from learning nonzero coefficients to features that have nothing to do with this class). Multiclass classification asymptotically succeeds when the survival dominates the contamination.

In Fig. 6, we plot experimental results using the bi-level ensemble model (Definition 4.2) for a setting where regression does not work but multiclass classification is conjectured to work. We plot quantities from Equation (22) in our error analysis. From subfigures (a),(b) and (c) we observe that while both survival and contamination are decreasing as we increase n, the survival/contamination ratio increases. The survival/contamination ratio growing with n is important for correct classification. The trend is very clear from the experimental results and indicates that continuing to grow n (together with the number of classes k, the number of features d, the number of favored features s, and the level of favoring as per our bi-level idealized model) would result in ever improving performance. Furthermore, we see that the empirical slope of these quantities on a log-log plot (and thus the power-law scaling of these quantities with respect to n) agree with the theoretical slopes calculated based on our conjecture.

Subfigure (d) plots the binary classification error when only trying to distinguish between the true class and one other particular class. (In this experiment, the true class was determined by feature 1. We calculate the binary error as the probability of misclassifiying a point from class 1 as belong to class 2 when we only compare the scores for class 1 and class 2.) We see that this error clearly decreases as we increase n. One way of thinking about successful multiclass classification is that the true class must win such pairwise competitions against all competing classes. Finally, subfigure (e) plots the total multiclass misclassification error overlaid with the number of classes. Here too we see a downward trend in classification error as n increases, even though we would have to go to significantly larger n than our compute could handle to see this error probability drop to very low values. Notice the integer effects arising from the number of classes k sometimes not growing with n that result in small upward spikes in the classification error.

J Comment on empirical eigenstructures of feature matrices

It is well known (for instance Remark 6 from Muthukumar et al. [56] that cites [78]) that for a spiked covariance model when the ratio of the top to the bottom eigenvalues grows as $\Omega(d/n)$, the top s eigenvalues can be estimated reliably from samples, even when the number of training samples n is less than the number of features d. The ratio of the top to the bottom eigenvalue in our bilevel model

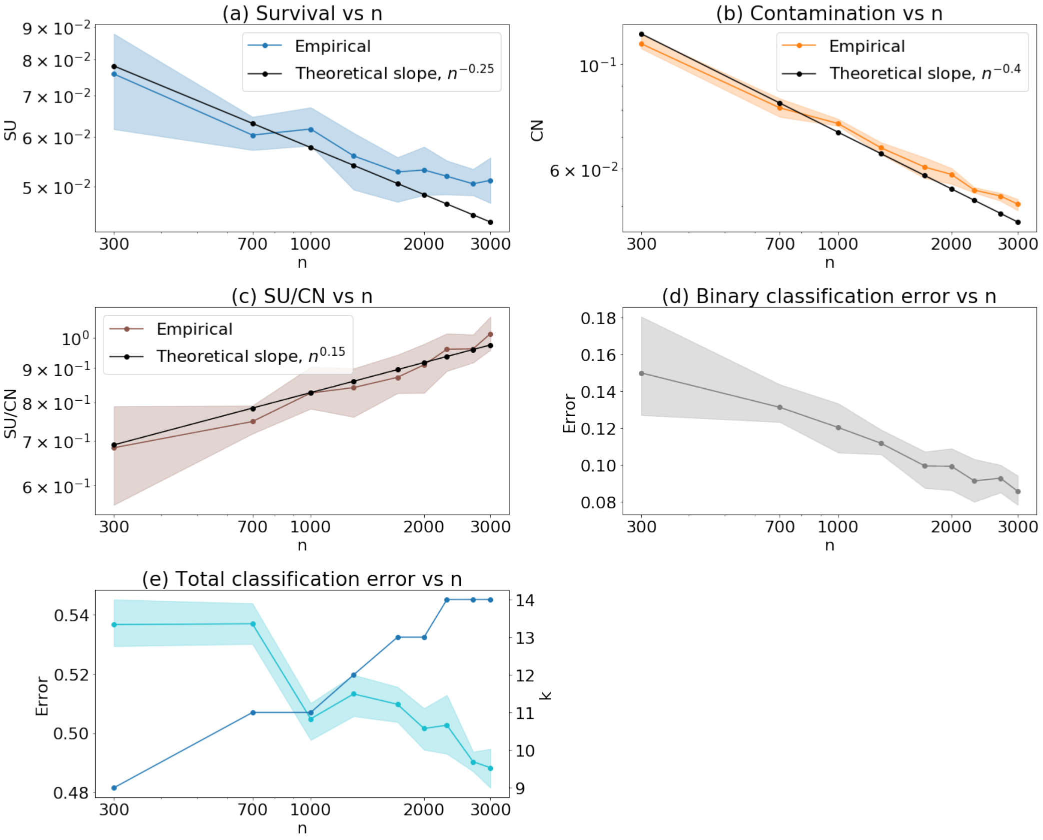


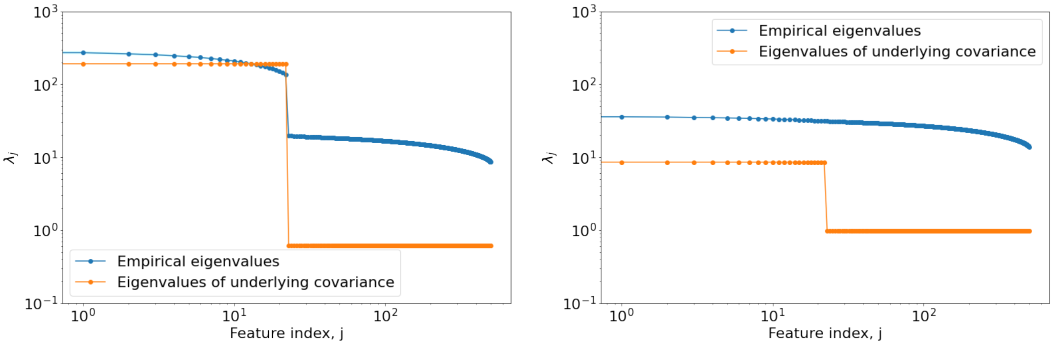
Figure 6: Experimental results using the bi-level ensemble model with $p = 1.5, q = 0.55, r = 0.5, t = 0.2$. Here, the number of training samples n varies from 300 to 3000 and the number of classes is computed as $k = \lfloor 3n^t \rfloor$ and varies from 9 to 14. We calculated the classification errors over a batch size of 10000, and ran 10 trials. The plots show the mean plotted with error bars corresponding to the $10^{th}, 90^{th}$ percentile values. We also plot the theoretical slopes for survival, contamination and the survival/contamination ratio based on our conjecture and notice that it closely matches the empirical slope of the quantities when plotted on a log-log scale. Notice that jaggedness in the plots is often due to integer effects as k grows or does not grow with n.

scales as

$$\frac{\frac{ad}{s}}{\frac{(1-a)d}{d-s}} = n^{p-q-r}, \tag{294}$$

and when $q+r < 1$, this ratio is larger than $d/n = n^{p-1}$. Fig. 7 shows empirical results of estimating the eigenvalues via the singular value decomposition of the training feature matrices. The visual distinction is quite striking. In the regime $q + r > 1$, the SVD of the training features matrix (and thus the empirical covariance matrix's eigenvalues) does not reveal that there are actually s favored features in the data. By contrast, in the regime $q + r < 1$, the SVD clearly shows an eigenvalue gap that reveals exactly what s is.

Theorem 5.1 shows that when $q + r > 1$ and there is not enough structure in the feature matrix to reliably estimate the top eigenvalues of the feature covariance matrix features, multiclass classification can still succeed for interpolating solutions as long as the number of classes does not increase too fast. It is because we are in the regime $q + r > 1$ that new techniques for analysis had to be developed in this paper, and the gap between the regime where we can prove the results and our conjectured results points interestingly to where there is a need for even better technique.



(a) $q = 0.15, q + r = 0.65$.

(b) $q = 0.65, q + r = 1.15$.

Figure 7: Estimating the eigenvalues of the covariance matrix of features empirically. Here $n = 400$ and the feature covariance structure follows the bi-level model with parameters $p = 1.5, r = 0.5$. Thus $d = 8001$ and $s = 21$ for both (a) and (b). The difference between (a) and (b) is in the level of favoring of favored features: with $q = 0.15$, (a) favors them more than (b) does with $q = 0.65$. In the regime where regression works, $q + r \leq 1$, we are able to accurately estimate the top s eigenvalues. In the regime where regression fails, $q + r > 1$, we are unable to estimate the top s eigenvalues accurately. The blue curve plots the estimated eigenvalues and the shaded region corresponds to the 10-90 percentile of the estimated values over 20 trials. Note, that there is only very small deviation across trials.

K A heuristic derivation of the main result

There is a significant amount technical work required in this paper to prove the main results, and Section F provides a conjecture for what the result should be, based on a more detailed look at our proof and where we suspect looseness. However, it is also possible to derive/conjecture the main result by another heuristic method in the style of Appendix B of Muthukumar et al. [55] and Appendix A of Muthukumar et al. [56]. At the heart of this heuristic method is an asymptotic linear-algebraic perspective driven by a few principles:

- When generalized linear models are sufficiently overparameterized, they tend to behave like the Fourier features model does on one-dimensional regularly spaced training data.
- If the favored features are approximately orthogonal on the training data, we can treat them as though they were exactly orthogonal on the training data as long as the number of favored features is sublinear in n, the number of training points.
- In the heavily overparameterized regime, for the bi-level model considered in this paper as well as in Muthukumar et al. [56] with s favored features, n training points, and d total features, min-norm-interpolation will behave as though there are $\frac{d}{n}$ exact aliases among the unfavored features for each of the favored features.
- Similarly, for this heavily overparameterized bi-level model, min-norm-interpolation will behave as though the unfavored features also provide another $n - s$ effectively orthogonal directions on the training data that also have $\frac{d}{n}$ exact aliases — all of these among the unfavored features.
- Any nonlinear function of a small group of training features can be viewed as having a linear part that is essentially captured by its projection onto that group of training features plus another component which is assumed to behave like "white noise" — equally spread across the n dimensions (assuming the size of the group is much smaller than n).
- On a test point, the different underlying features behave like uncorrelated random variables whose means are 0 and whose variances are proportional to the magnitude-squared of their coefficient times their inherent weighting squared.

These heuristics, together with other standard approximations like $1 + |x| \approx |x|$ when $|x| \gg 1$ and $1 + |x| \approx 1$ when $|x| \ll 1$ permit relatively simple calculations to be used together with the survival/contamination style of analysis to predict when regression and classification type problems will succeed. This is done in Appendix B of Muthukumar et al. [56] for binary classification. Here, it is useful to recap the key insights from that paper in the context of the same bi-level model used here:

- When the underlying score function that we would like to learn is a linear combination of the favored features, the survival captures the extent to which that particular linear combination is present in the learned score. Without loss of generality, because all favored features are equally favored, it suffices to consider a desired score that is purely one of the favored features, normalized to have unit variance. In this case, the coefficient learned for that (normalized) feature represents the survival.
- The contamination captures everything else that is learned — all the learned coefficients of features other than the true feature — and is measured in terms of the standard deviation of the predictions due to those (falsely) learned coefficients.
- For learning in a regression problem to asymptotically generalize, the survival must tend to 1 and the contamination must tend to 0. This is because the score function itself must be learned in a mean-square-error sense.
- For learning in a binary classification problem to asymptotically generalize, all we need to get right is the sign of the learned score. To get the sign right, it suffices to have the ratio of survival to contamination go to infinity. This can happen even if both the survival and contamination tend to 0, as long as they do so at the right rates relative to each other.

K.1 Understanding what we learn from multiclass training data

Using the above heuristics and reusing calculations already done in Appendix A of Muthukumar et al. [56], we can see how survival and contamination asymptotically behave for multi-class training data. For class m, we consider the score function we learn by interpolating the zero-mean one-hot encoding for the n training points as represented by $\mathbf{y}_m$. Within $\mathbf{y}_m$, we are going to assume that there are exactly $\frac{n}{k}$ positive examples for the m-th class, and each of those is represented with a $1 - \frac{1}{k}$ in the appropriate position of $\mathbf{y}_m$. There are also $n - \frac{n}{k}$ negative examples for the m-th class, each represented with a $-\frac{1}{k}$ in the appropriate position.

The total "energy" (norm-squared) in this vector $\mathbf{y}_m$ is thus

$$\begin{aligned}\frac{n}{k}\left(1-\frac{1}{k}\right)^2 + \left(n - \frac{n}{k}\right)\frac{1}{k^2} &= \frac{n}{k} - \frac{n}{k^2} \\ &\approx \frac{n}{k}\end{aligned}$$

where the final approximation is due to k asymptotically growing as n^t to infinity.

We can understand the process of min-norm interpolation as proceeding in two conceptual steps. First, this n-dimensional vector $\mathbf{y}_m$ is decomposed into the s orthogonal directions represented by the s favored features and the $n-s$ unfavored synthetic directions according to the heuristic above. For this step, the level of favoring does not matter. After that, the level of favoring determines precisely how each of those directions is split across the representative favored feature (if any) and its $\frac{d}{n}$ unfavored aliases.

K.1.1 Survival

How much of $\mathbf{y}_m$ ends up going into the direction representing the true favored feature m? Because of orthogonality, we can simply look at the correlation between $\mathbf{y}_m$ and a normalized vector in the direction of $\mathbf{z}_m$. We can approximate the standard Gaussian in $\mathbf{z}_m$ which wins a competition with $k-1$ other iid standard Gaussians as being a constant $\sqrt{\ln k}$, which as compared to the polynomial scalings relevant here, might as well just be the constant 1. With that, we can consider all the other entries of $\mathbf{z}_m$ as being basically their mean, which by the same logic is essentially $\frac{-1}{k}$. This gives a total correlation of

$$\frac{1}{\sqrt{n}}\left(\frac{n}{k}\cdot\left(1-\frac{1}{k}\right)\cdot 1 + \left(n-\frac{n}{k}\right)\left(-\frac{1}{k}\right)\left(-\frac{1}{k}\right)\right) = \frac{1}{\sqrt{n}}\left(\frac{n}{k} - \frac{n}{k^3}\right) \tag{295}$$

$$\approx \frac{\sqrt{n}}{k} \tag{296}$$

with a normalized vector $\frac{\mathbf{z}_m}{\|\mathbf{z}_m\|}$.

Notice that if we had done this correlation with $\mathbf{z}_m$ itself instead of $\mathbf{y}_m$, we would have gotten simply $\sqrt{n}$. This shows how multiclass training data immediately reduces the survival by an asymptotic factor of k relative to noise-free regression training data.

K.1.2 Contamination

To understand contamination, we simply break it down into two sources: the other $s - 1 \approx s$ favored features and the unfavored features.

Here, we leverage the heuristic that the total $\frac{n}{k}$ energy of the training labels $\mathbf{y}_m$ has to be split across the true feature m, the other k label-defining features that are not the true feature, the other $s - k \approx s$ favored features, and the unfavored features.

For positive training examples for class m, the other label-defining features have a mean value of $-\frac{1}{k}$ since they are not the max. For negative training examples of class m, the mean value for other label-defining features is essentially zero. The consequence of this (by a calculation exactly analogous to (296)) is that the projected correlation for these is like $-\frac{\sqrt{n}}{k^2}$ which might as well be zero.

How much energy is left in $\mathbf{y}_m$ after we remove the components along the k label-defining directions that could have any linear relationship to it?

$$\frac{n}{k} - \left(\frac{\sqrt{n}}{k}\right)^2 - (k-1)\left(-\frac{\sqrt{n}}{k^2}\right)^2 \approx \frac{n}{k} - \frac{n}{k^2} - \frac{n}{k^3} \tag{297}$$

$$\approx \frac{n}{k}. \tag{298}$$

Asymptotically, all the energy is still left. This can now be divided equally across the $n - k \approx n$ orthogonal directions by the heuristic.

This means that a fraction $\frac{s}{n}$ of this ends up as contamination in the favored features, which is a total energy of $\frac{s}{n} \cdot \frac{n}{k} = \frac{s}{k}$.

Meanwhile, there are $n - s$ other directions that are only represented by the unfavored features. This has total energy $\frac{n-s}{n} \cdot \frac{n}{k} = \frac{n-s}{k} \approx \frac{n}{k}$.

K.1.3 When does survival dominate contamination?

There are qualitatively two kinds of contamination: coming from favored features and coming from unfavored features.

For favored features (like the true feature), there is a further split that happens as the min-norm-interpolation solution splits the coefficients themselves across the favored feature and its effectively $\frac{d}{n} = n^{p-1}$ unfavored aliases. Adapting the notation in Appendix A of Muthukumar et al. [56], let's say a fraction α goes onto the favored feature itself. Equation (23) in Muthukumar et al. [56] tells us that

$$\alpha \approx \begin{cases} 1 & \text{if } q < 1 - r \\ n^{-(q-(1-r))} & \text{if } q > 1 - r \end{cases}. \tag{299}$$

Consequently, the actually survived signal scales as $\alpha\frac{\sqrt{n}}{k}$ while in those same units, the standard deviation of favored contamination scales as $\alpha\sqrt{\frac{s}{k}}$. From this we immediately see that the α cancel and the condition for classification working (as far as the favored features are concerned) is that

$$\frac{\sqrt{n}}{k} \gg \sqrt{\frac{s}{k}} \tag{300}$$

which implies

$$k \ll \frac{n}{s} \tag{301}$$

and since $k = n^t$ and $s = n^r$, this gives us a condition that $t < 1 - r$ for survival to dominate contamination from favored features.

This $t < 1 - r$ condition turns out not to care about whether we are in the regime where regression works (i.e. $q < 1 - r$) or not.

To understand the impact of contamination due to unfavored features, it is important to recall that all of that gets split across the $\frac{d}{n}$ aliases in our heuristic calculation and so the standard deviation of the contamination due to unfavored features scales as $\sqrt{\frac{n}{k} \cdot \frac{n}{d}} = \frac{n}{\sqrt{kd}} = n^{1-\frac{p}{2}-\frac{t}{2}}$. Meanwhile, survival behaves as $\alpha \frac{\sqrt{n}}{k} = \alpha n^{\frac{1}{2}-t}$.

For survival to dominate contamination from unfavored features, we need:

$$\alpha n^{\frac{1}{2}-t} \gg n^{1-\frac{p}{2}-\frac{t}{2}} \tag{302}$$

$$\alpha \gg n^{\frac{t-(p-1)}{2}} \tag{303}$$

$$\alpha n^{\frac{p-1}{2}} \gg n^{\frac{t}{2}} \tag{304}$$

Substituting in for the case $q < 1 - r$ from (299), we immediately see the condition $t < p - 1$ for survival to dominate contamination from unfavored features.

For the case $q > 1 - r$, the substitution of $n^{-(q-(1-r))}$ for α in (304) gives us:

$$\frac{t}{2} < \frac{p-1}{2} - (q - (1-r)) \tag{305}$$

$$t < (p-1) - 2(q - (1-r)). \tag{306}$$

Notice that at the boundary $q = (1 - r)$, the condition (306) matches the $t < p - 1$ condition for the region $q < 1 - r$.

Finally, note that our model requires $t < r$ so that the features that determine class labels are themselves favored. Putting all the terms together, we get the condition:

$$t < \min(r, 1-r, (p-1) - 2 \cdot \max(0, q - (1-r))) \tag{307}$$

which agrees with the overall conjectured condition in (23).

K.2 Why is this sufficient for multiclass classification

Notice the heuristic derivation above is actually just saying that binary classification will succeed (i.e. we will typically have the m-th learned score agree in sign with the actual m-th feature for a test point) in the given regime, assuming we trained with multiclass training data.

But successful generalization for multiclass classification requires more — it requires the true class's score to win a competition against the competing classes.

Here, we can simply leverage a quick heuristic calculation involving order statistics. We know that the max of k i.i.d. random variables will concentrate to a neighborhood of where the Complementary Cumulative Distribution Function (CCDF) of the underlying random variable reaches $\frac{1}{k}$. Meanwhile, the second biggest of those variables will be around where CCDF hits $\frac{2}{k}$. For a standard Gaussian, these are basically at $\sqrt{2 \ln k}$ and then at $\sqrt{2(\ln k - \ln 2)}$. The gap between these is

$$\sqrt{2 \ln k} - \sqrt{2 \ln k - 2 \ln 2} \approx \frac{1}{2\sqrt{2 \ln k}} 2 \ln 2 \tag{308}$$

$$= \frac{\ln 2}{\sqrt{2 \ln k}} \tag{309}$$

and so what matters is that the contamination relative to the survival needs to go to zero faster than $\frac{1}{\sqrt{\ln k}} = \frac{1}{\sqrt{t \ln n}}$. However, all the scalings here are polynomial in n and so if the contamination is asymptotically smaller than survival, it is smaller by a factor that is polynomial in n. This will beat the $\frac{1}{\sqrt{\ln n}}$ scaling of this gap, resulting in successful multiclass classification.

Notice that we crucially used the Gaussian nature of the underlying features[19] here. If the underlying features were uniformly distributed on $[-1, 1]$ instead, then the gap between the biggest and second biggest feature would instead decay polynomially like $\frac{1}{k}$. In that case, the relative contamination would have to go to zero fast enough to overcome this, making the conjectured multiclass region different.

[19] Any sufficiently thick tail would do. For example, an exponential or Laplacian feature would also work here.

K.3 Comments on the power of the heuristic calculation

The heuristic calculation being done here operated at a level of abstraction for which the fine details of the model do not matter. The total energy in the $\mathbf{y}_m$ just comes from the nature of multiclass training data where every point is labeled with only one class as its label and we have a balanced training set. The dominant contamination analysis follows simply from that.

All we need to verify is the survival analysis for alternative models. Here, it is also immediately clear that the asymptotic survival calculation would work just as well if one used a Gaussian mixture model with orthogonal class means or a logistic-softmax model with orthogonal classes — like the models in Wang et al. [77].

In fact, based on this heuristic calculation, we can conjecture that even if the training data were less informative (e.g. the training label did not actually correspond to the class whose feature had the true max but instead, just one of the positive feature classes was chosen uniformly at random as the training label), multiclass classification would still asymptotically succeed in the conjectured regimes in terms of reliably predicting the class whose feature is biggest. What matters is just that positive examples of a class show a significant amount of "classiness" on average in their training features.

Note however that the heuristic calculation here is just that — a heuristic calculation in what some might call the "physics style." Proving the conjectured region is in fact correct requires us to leverage the specific details of the models in our proofs.