Lower Bounds for Multiclass Classification with Overparameterized Linear Models

Anonymous Authors

***Abstract*—Subramanian *et al.* [1] introduced an asymptotic Gaussian-features model for overparameterized multiclass classification in which the number of classes, training points, and parameters all go to infinity. They provided some achievable regions where min-norm interpolating classifiers successfully asymptotically generalize as well as conjecturing the full form of the region based on a heuristic analysis. Here, we introduce a converse for such min-norm interpolating classifiers in their model which fully matches their conjectured regions. The key technical tool is a variant of the Hanson-Wright concentration inequality that applies to the sparse bilinear forms that arise.**

I. INTRODUCTION

Classical statistical learning theory intuition predicts that highly expressive models, which can interpolate random labels [2], [3], ought not to generalize well. However, deep learning practice has seen such models performing well when trained with good labels. Resolving this apparent contradiction has recently been the focus of a multitude of works, and this paper builds on one particular thread of investigation. To be self-contained, this introduction will quickly summarize this thread but we direct the reader to [1] and the references cited therein since space here is very constrained. A broader picture that encompasses other threads can be found in [4]–[6].

Our thread begins with a recent line of work that analyzes the generalization behavior of overparameterized linear models for regression [7]–[11]. These simple models demonstrate how the capacity to interpolate noise can actually aid in generalization: training noise can be harmlessly absorbed by the overparameterized model without contaminating predictions on test points. In effect, extra features can be regularizing (in the context of descent algorithms' implicit regularization [12]–[15]), but an excessive amount of such regularization causes regression to fail because even the true signal will not survive the training process.

The thread continues in a line of work that studies binary classification [16]–[18] in similar overparameterized linear models. While confirming that the basic story is similar to regression, these works identify a further surprise: binary classification can work in some regimes where the corresponding regression problem would not work[1] due to the regularizing effect of overparameterization being too strong. Just as in the regression case, the results here are sharp in toy models: we can exactly characterize where binary classification using an interpolating classifier asymptotically generalizes.

With binary classification better understood, the thread continues to multiclass classification. After all, the current wave of deep learning enthusiasm originated in breakthrough performance in multiclass classification, and we have seen a decade of ever larger networks trained on ever larger datasets with ever more classes [20]. Using similar toy models [11], [16], [21], the constant number of classes case was studied in [22] to recover results similar to binary classification. Subramanian *et al.* [1] further introduced a model where the number of classes grows with the number of training points and proved an achievability result on how fast the number of classes can grow while still allowing the interpolating classifier to asymptotically generalize. While [1] gave a conjecture for what the full region should be, there was no converse proof.

In this paper, we prove a weak converse that exactly matches what was conjectured in [1]. To do so, we leverage a new tool: a variant of the Hanson-Wright concentration inequality that applies to bilinear forms and takes advantage of the "sparsity" inherent in multiclass classification.

II. PROBLEM SETUP

We consider the multiclass classification problem with k classes. The following exposition is lifted from [1], but we include it for the sake of being self-contained. The training data consists of n pairs $\{\boldsymbol{x}_i, \ell_i\}_{i=1}^n$ where $\boldsymbol{x}_i \in \mathbb{R}^d$ are i.i.d standard Gaussian vectors[2]. We assume that the labels $\ell_i \in [k]$ are generated as follows.

Assumption 1 (1-sparse noiseless model): The class labels ℓ_i are generated based on which of the first k dimensions of a point $\boldsymbol{x}_i$ has the largest value,

$$\ell_i = \arg\max_{m\in[k]} \boldsymbol{x}_i[m]. \tag{1}$$

For a vector $\boldsymbol{x}$, we index its jth entry with $\boldsymbol{x}[j]$. Hence, under Assumption 1, $\boldsymbol{x}_i[m]$ can be interpreted as how representative of class m the ith training point is.

For clarity of exposition, we make explicit a feature weighting that transforms the training points as follows:

$$\boldsymbol{x}_i^w[j] = \sqrt{\lambda_j}\boldsymbol{x}_i[j] \quad \forall j \in [d]. \tag{2}$$

---

[1] The phenomenon of regression failing in the overparameterized regime is inextricably linked to the empirical covariance of the data not revealing the spiked reality of the underlying true covariance [19]. See Appendix J of [1].

[2] Following previous work, we are staying within a Gaussian features framework. However, recent developments have confirmed that these models are actually predictive when the features arise from nonlinearities in a lifting, as long as there is enough randomness underneath [23]–[27].

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.[3]

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

$$\boldsymbol{X}^w = \begin{bmatrix} \boldsymbol{x}_1^w & \cdots & \boldsymbol{x}_n^w \end{bmatrix}^\top = \begin{bmatrix} \sqrt{\lambda_1}\boldsymbol{z}_1 & \cdots & \sqrt{\lambda_d}\boldsymbol{z}_d \end{bmatrix} \quad (3)$$

where we introduce the notation $\boldsymbol{z}_j \in \mathbb{R}^n$ to contain the j^{th} feature from the n training points. Note that $\boldsymbol{z}_j \sim N(0, \boldsymbol{I}_n)$ are i.i.d Gaussians. We use a one-hot encoding for representing the labels as the matrix $\boldsymbol{Y}^{\mathsf{oh}} \in \mathbb{R}^{n\times k}$

$$\boldsymbol{Y}^{\mathsf{oh}} = \begin{bmatrix} \boldsymbol{y}_1^{\mathsf{oh}} & \cdots & \boldsymbol{y}_k^{\mathsf{oh}} \end{bmatrix}, \quad (4)$$

where

$$\boldsymbol{y}_m^{\mathsf{oh}}[i] = \begin{cases} 1, & \text{if } \ell_i = m \\ 0, & \text{otherwise} \end{cases}. \quad (5)$$

Since we consider linear models, we center the one-hot encodings by subtracting $\frac{1}{k}$ from each entry, and define

$$\boldsymbol{y}_m \triangleq \boldsymbol{y}_m^{\mathsf{oh}} - \frac{1}{k}\mathbf{1}. \quad (6)$$

Our classifier consists of k coefficient vectors $\widehat{\boldsymbol{f}}_m$ for $m \in [k]$ that are learned by minimum-norm interpolation of the zero-mean one-hot variants using the weighted features:[4]

$$\widehat{\boldsymbol{f}}_m = \arg\min_{\boldsymbol{f}} \|\boldsymbol{f}\|_2 \quad (7)$$

$$\text{s.t. } \boldsymbol{X}^w\boldsymbol{f} = \boldsymbol{y}_m. \quad (8)$$

We can express these coefficients in closed form as

$$\widehat{\boldsymbol{f}}_m = (\boldsymbol{X}^w)^\top \left(\boldsymbol{X}^w(\boldsymbol{X}^w)^\top\right)^{-1}\boldsymbol{y}_m. \quad (9)$$

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

$$\hat{\ell} = \arg\max_{1\le m\le k} \widehat{\boldsymbol{f}}_m^\top \boldsymbol{x}_{\mathsf{test}}^w. \quad (10)$$

By assumption, a misclassification event $\mathcal{E}_{\mathsf{err}}$ occurs whenever

$$\arg\max_{1\le m\le k} \boldsymbol{x}_{\mathsf{test}}[m] \neq \arg\max_{1\le m\le k} \widehat{\boldsymbol{f}}_m^\top \boldsymbol{x}_{\mathsf{test}}^w. \quad (11)$$

In this paper, we prove a converse result to the positive result for generalization in [1]. In particular, we determine sufficient conditions under which the probability of misclassification (computed over the randomness in both the training data and test point) is *bounded below by a positive constant* in an asymptotic regime where the number of training points, features, classes, and feature weights all scale according to the bi-level ensemble model, which we now formally define.

Definition 1 (Bi-level ensemble): The bi-level ensemble is parameterized by p, q, r and t where $p > 1$, $0 \le r < 1$, $0 < q < (p - r)$ and $0 \le t < r$. 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) all scale with the number of training points (n) as follows:

$$d = \lfloor n^p \rfloor, s = \lfloor n^r \rfloor, a = n^{-q}, k = c_k\lfloor n^t \rfloor, \quad (12)$$

where c_k is a positive integer. Furthermore, the feature weights ($\sqrt{\lambda_j}$) scale according to the following definition:

$$\sqrt{\lambda_j} = \begin{cases} \sqrt{\frac{ad}{s}}, & 1 \le j \le s \\ \sqrt{\frac{(1-a)d}{d-s}}, & \text{otherwise} \end{cases}. \quad (13)$$

We introduce the notation $\lambda_F \triangleq \frac{ad}{s}$ and $\lambda_U \triangleq \frac{(1-a)d}{d-s}$ to distinguish between the (squared) favored and unfavored weights, respectively.

We provide a visualization of the bi-level model in Figure 1, reproduced from [1].

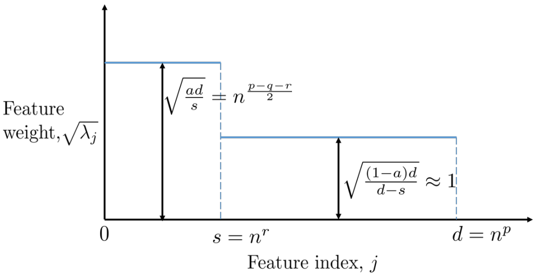
Feature
weight, $\sqrt{\lambda_j}$
$\sqrt{\frac{ad}{s}} = n^{\frac{p-q-r}{2}}$
$\sqrt{\frac{(1-a)d}{d-s}} \approx 1$
0
$s = n^r$
$d = n^p$
Feature index, j

Fig. 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 Σ in a Gaussian model for the covariates as in [9].

Subramanian *et al.* [1] use heuristic calculations to conjecture necessary and sufficient conditions for the bi-level model to generalize. We restate it here, as it provides the primary benchmark for our theoretical result.

Conjecture 1 (Conjectured bi-level regions): Under the bi-level ensemble model 1, when the true data generating process is 1-sparse (Assumption 1), as $n \to \infty$, in the regime where $q + r > 1$, the probability of misclassification $\mathbb{P}[\mathcal{E}_{\mathsf{err}}]$ satisfies[5]

$$\mathbb{P}[\mathcal{E}_{\mathsf{err}}] \to \begin{cases} 0, & \text{if } t < \min\{1 - r, p + 1 - 2(q + r)\} \\ 1, & \text{if } t > \min\{1 - r, p + 1 - 2(q + r)\} \end{cases}. \quad (14)$$

III. MAIN RESULT: WHERE MIN-NORM INTERPOLATION IN THE BI-LEVEL MODEL FAILS TO GENERALIZE

For the sake of comparison to [1], we only consider the regime where regression provably fails, i.e. $q + r > 1$ (see e.g. [9], [16]).

---

[3] Our weighted feature model is equivalent to the one used in other works (e.g. [16]) that assume that the covariates come from a $d-$dimensional anisotropic Gaussian with a covariance matrix Σ that favors the truly important directions. These directions do not have to be axis-aligned — we make that assumption only for notational convenience. In reality, the optimizer will never know these directions *a priori*.

[4] The classifier learned via this method is equivalent to those obtained by other natural training methods under sufficient overparameterization [22].

[5] We have omitted the constraint $t < r$ as this is included in the definition of the bi-level model.

Theorem 1 (Impossibility for bi-level model): Under the bi-level ensemble model 1, when the true data generating process is 1-sparse (Assumption 1), the probability of misclassification $\mathbb{P}[\mathcal{E}_{\mathsf{err}}] \geq \frac{1}{2}$ as $n \to \infty$ if the following conditions hold:

$$t > \min\{1-r, p+1-2(q+r)\} \tag{15}$$
$$q+r > 1. \tag{16}$$

We now quote the corresponding positive result for the bi-level model, which is Theorem 5.1 in [1].

Theorem 2 (Generalization for bi-level model): Under the bi-level ensemble model 1, when the true data generating process is 1-sparse (Assumption 1), the probability of misclassification $\mathbb{P}[\mathcal{E}_{\mathsf{err}}] \to 0$ as $n \to \infty$ if the following hold:

$$t < \min\{1-r, p+1-2(q+r), p-2, 2q+r-2\} \tag{17}$$
$$q+r > 1. \tag{18}$$

As alluded to in the introduction, our main result Theorem 1 is a weak converse result. With some extra effort, we believe this can be strengthened for a strong converse with asymptotic probability of misclassification of $1-\frac{1}{k}$, which would match random guessing. Our converse result Theorem 1 fully resolves the impossibility side of Conjecture 1 in the regime where regression fails. We expect our techniques to carry over to the $q+r<1$ regime where regression works, but we omit these details due to space constraints.

In Figure 2, we compare Theorems 1 and 2 by visualizing the regimes where they hold; this figure parallels Fig 2. in [1]. The figure depicts slices of the four dimensional scaling parameter space of $p, q, r,$ and t. We fix the value of q to 0.75, as our result fully matches the conjectured regimes. Note that there is looseness in the positive result in [1]. We expect our tighter analysis techniques from this paper to close that gap, but we omit these details for the sake of space. We point out that in (2b), the boundary of the region where multiclass classification fails contains two slopes. These slopes arise from the two conditions in Theorem 1.

IV. PROOF SKETCH AND MAIN TECHNIQUES

In this section we briefly describe the high level ideas of the proof of Theorem 1. Parallel to [1], the starting point is writing out a sufficient (instead of necessary) condition for misclassification. From there, the crux is proving the correct order of growth of a certain signal-to-noise ratio, which we will define later.

Assume without loss of generality that the test point $\boldsymbol{x}_{\mathsf{test}} \sim N(0, \boldsymbol{I}_d)$ has true label α for some $\alpha \in [k]$. Let $\boldsymbol{x}^w_{\mathsf{test}}$ be the weighted version of this test point. From (11), an equivalent condition for misclassification is that for some $\beta \neq \alpha, \beta \in [k]$, we have $\widehat{\boldsymbol{f}}_\alpha^\top \boldsymbol{x}^w_{\mathsf{test}} < \widehat{\boldsymbol{f}}_\beta^\top \boldsymbol{x}^w_{\mathsf{test}}$, i.e. the score for β outcompetes the score for α. Define the Gram matrix $\boldsymbol{A} \triangleq \boldsymbol{X}^w(\boldsymbol{X}^w)^\top$, the relative label vector $\boldsymbol{\Delta} y \triangleq \boldsymbol{y}_\alpha - \boldsymbol{y}_\beta \in \{-1,0,1\}^n$, and the relative survival vector $\widehat{\boldsymbol{h}}_{\alpha,\beta} \in \mathbb{R}^d$ which compares the signal from α and β:

$$\widehat{\boldsymbol{h}}_{\alpha,\beta}[j] \triangleq \lambda_j^{-1/2}(\widehat{\boldsymbol{f}}_\alpha[j] - \widehat{\boldsymbol{f}}_\beta[j]) \tag{19}$$
$$= \boldsymbol{z}_j^\top \boldsymbol{A}^{-1} \boldsymbol{\Delta} y, \tag{20}$$

where to obtain the last line we have used (9). By converting the misclassification condition into the unweighted feature space we see that we will have errors when

$$\begin{aligned} &\lambda_\alpha \widehat{\boldsymbol{h}}_{\alpha,\beta}[\alpha]\boldsymbol{x}_{\mathsf{test}}[\alpha] - \lambda_\beta \widehat{\boldsymbol{h}}_{\beta,\alpha}[\beta]\boldsymbol{x}_{\mathsf{test}}[\beta] \\ &\quad < \sum_{j \notin \{\alpha,\beta\}} \lambda_j \widehat{\boldsymbol{h}}_{\beta,\alpha}[j]\boldsymbol{x}_{\mathsf{test}}[j]. \end{aligned} \tag{21}$$

Define the contamination term $\mathsf{CN}_{\alpha,\beta}$:

$$\mathsf{CN}_{\alpha,\beta} \triangleq \sqrt{\sum_{j \notin \{\alpha,\beta\}} \lambda_j^2 (\widehat{\boldsymbol{h}}_{\beta,\alpha}[j])^2}. \tag{22}$$

Note that $\mathsf{CN}_{\alpha,\beta}$ normalizes the RHS of (21) into a standard Gaussian. Indeed, define

$$Z^{(\beta)} \triangleq \frac{1}{\mathsf{CN}_{\alpha,\beta}} \sum_{j \notin \{\alpha,\beta\}} \lambda_j \widehat{\boldsymbol{h}}_{\beta,\alpha}[j]\boldsymbol{x}_{\mathsf{test}}[j] \sim N(0,1). \tag{23}$$

Since $\alpha, \beta \in [k]$ are favored, we have $\lambda_\alpha = \lambda_\beta = \lambda_F$. Taking an absolute value of the LHS of (21), a sufficient condition for misclassification is

$$\frac{\lambda_F}{\mathsf{CN}_{\alpha,\beta}}\left(|\boldsymbol{x}_{\mathsf{test}}[\alpha]|\left|\widehat{\boldsymbol{h}}_{\alpha,\beta}[\alpha]\right| + |\boldsymbol{x}_{\mathsf{test}}[\beta]|\left|\widehat{\boldsymbol{h}}_{\beta,\alpha}[\beta]\right|\right) < Z^{(\beta)}. \tag{24}$$

By standard subgaussian maximal inequalities, $|\boldsymbol{x}_{\mathsf{test}}[i]| = O(\sqrt{\log(nk)})$ with high probability. Hence, misclassification occurs with nonvanishing probability if the survival to contamination ratio $\lambda_F|\widehat{\boldsymbol{h}}_{\alpha,\beta}[\alpha]|/\mathsf{CN}_{\alpha,\beta} \leq n^{-w}$ for some $w > 0$, and similarly for $\lambda_F|\widehat{\boldsymbol{h}}_{\beta,\alpha}[\beta]|/\mathsf{CN}_{\alpha,\beta}$. Taking this result on faith for now, we can deduce that since $N(0,1)$ is a symmetric continuous distribution, some $Z^{(\beta)}$ outcompetes with probability at least $\frac{1}{2} - o(1)$.

We now discuss the high level proof ideas for proving that the survival to contamination ratio shrinks. We highlight where our techniques differ from those of Subramanian *et al.* [1]. To understand the relative survival and contamination, we must understand the bilinear forms $\widehat{\boldsymbol{h}}_{\alpha,\beta}[j] = \boldsymbol{z}_j^\top \boldsymbol{A}^{-1}\boldsymbol{\Delta} y$. The main source of inspiration for bounding these bilinear forms is the heuristic style of calculation that leads to Conjecture 1.

In the regime where regression fails, $\boldsymbol{A}^{-1}$ turns out to have a *flat* spectrum. In other words, $\boldsymbol{A}^{-1}$ is very close to a scaled identity matrix. If we assume that $\boldsymbol{A}^{-1}$ is *exactly* equal to a scaled identity matrix, then the survival for α is proportional to $\boldsymbol{z}_\alpha^\top \boldsymbol{\Delta} y$. This is an inner product between two random vectors. We point out that it is crucial that $\boldsymbol{\Delta} y$ is a sparse vector; it only has $2n/k$ nonzero entries in expectation. A quick computation reveals that $\mathbb{E}[\boldsymbol{z}_\alpha^\top \boldsymbol{\Delta} y] \approx O(\frac{n}{k})$. With some additional effort, one can show that the inner product concentrates. The same argument applies verbatim for β.

Similarly, to lower bound the contamination terms we lower bound $\left|\boldsymbol{z}_j^\top \boldsymbol{\Delta} y\right|$ for $j \notin \{\alpha,\beta\}$. By symmetry $\mathbb{E}[\boldsymbol{z}_j^\top \boldsymbol{\Delta} y] = 0$. The growth of the contamination term is therefore determined by the concentration radius, which is where we hope to leverage sparsity. This is the heart of Theorem 4, which may be of independent interest; we present it in the following section.

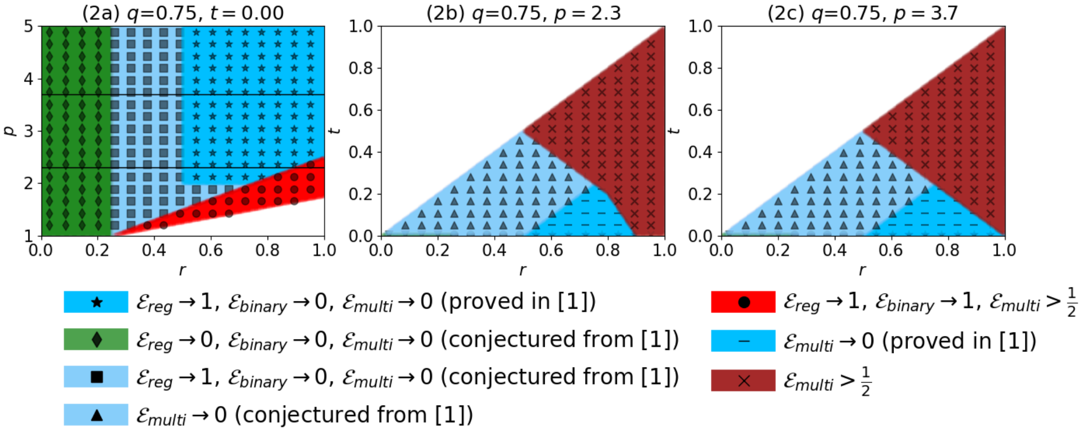

Fig. 2. Example of regimes for multiclass/binary classification and regression. The white regions correspond to invalid regimes under the bi-level model.

A. *On the role of Hanson-Wright concentration*

In reality, $\boldsymbol{A}^{-1}$ is not actually perfectly flat, so we cannot immediately reduce the bilinear form $\boldsymbol{z}_j^\top \boldsymbol{A}^{-1}\boldsymbol{\Delta} y$ to a simple inner product. Instead, we turn to the well-known Hanson-Wright inequality [28], which tells us that quadratic forms of random vectors with independent, mean zero, subgaussian entries concentrate around their mean. It was used extensively to study binary classification [16] and the positive result for multiclass classification [1] in the bi-level model. However, there are several key differences in our use of Hanson-Wright. At a high level, our analysis is optimal because it cleverly exploits the dependence structure in the problem and fully leverages a new variant of Hanson-Wright which explicitly uses the sparsity inherent to the multiclass problem.

For reference, we state the original version of the Hanson-Wright inequality proved in [28]. First, for the sake of precision, we define the subgaussian norm $\|\xi\|_{\psi_2}$ [29] as

$$\|\xi\|_{\psi_2} = \inf_{K>0} \{K : \mathbb{E}\exp(\xi^2/K^2) \le 2\}, \tag{25}$$

Theorem 3 (Hanson-Wright for quadratic forms, from [28]): Let $\boldsymbol{x} \in \mathbb{R}^n$ be a random vector composed of independent random variables that are zero mean and have subgaussian norm at most K. There exists universal constant $c > 0$ such that for any deterministic $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ and $\epsilon \ge 0$,

$$\begin{aligned}\mathbb{P}[|\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x}]| > \epsilon] \\ \le 2\exp\left(-c\min\left\{\frac{\epsilon^2}{K^4\|\boldsymbol{M}\|_F^2}, \frac{\epsilon}{K^2\|\boldsymbol{M}\|_2}\right\}\right).\end{aligned}$$

The original Hanson-Wright inequality quoted above only applies to quadratic forms, and moreover assumes that the matrix $\boldsymbol{M}$ is deterministic. In our setting, we have a bilinear form $\boldsymbol{z}_j^\top \boldsymbol{A}^{-1}\boldsymbol{\Delta} y$, where $\boldsymbol{A}^{-1}$ is random. One can condition on the realization of $\boldsymbol{A}^{-1}$, but this removes independence and alters the distributions of the random variables involved. Of course, if we condition on a random matrix which is independent of the random vectors, then there is no issue.

Assuming a way around the independence issue, one could decompose the bilinear form with the identity

$$\begin{aligned}4\boldsymbol{z}_j^\top \boldsymbol{A}^{-1}\boldsymbol{\Delta} y &= (\boldsymbol{z}_j + \boldsymbol{\Delta} y)^\top \boldsymbol{A}^{-1}(\boldsymbol{z}_j + \boldsymbol{\Delta} y) &(26)\\ &\quad - (\boldsymbol{z}_j - \boldsymbol{\Delta} y)^\top \boldsymbol{A}^{-1}(\boldsymbol{z}_j - \boldsymbol{\Delta} y). &(27)\end{aligned}$$

This trick is used in both [1], [16]. In the binary classification case, one regains complete independence by using a leave-one-out trick. More precisely, define the leave-one-out matrix $\boldsymbol{A}_{-j} = \sum_{i\neq j} \lambda_i \boldsymbol{z}_i \boldsymbol{z}_i^\top$ and let $\boldsymbol{y} \in \{\pm 1\}^n$ be the binary label vector. Then $\boldsymbol{A}_{-j}$ is evidently independent of $\boldsymbol{z}_j$ and $\boldsymbol{y}$, and the Sherman-Morrison formula implies that $\boldsymbol{z}_j^\top \boldsymbol{A}^{-1}\boldsymbol{y} = \frac{\boldsymbol{z}_j^\top \boldsymbol{A}_{-j}^{-1}\boldsymbol{y}}{1+\boldsymbol{z}_j^\top \boldsymbol{A}_{-j}^{-1}\boldsymbol{z}_j}$. Because the denominator is a scalar which concentrates well due to Hanson-Wright, this allows for a completely tight characterization of binary classification.

However, this trick does not immediately work in the multiclass setting, because the labels depend on all of the $k > 1$ label-defining features. Here, one needs to potentially remove $\omega(1)$ features from the Gram matrix $\boldsymbol{A}$ to regain independence. In [1], they eschew Sherman-Morrison entirely and directly exploit the fact that $\boldsymbol{A}^{-1}$ is flat (as $q + r > 1$). More precisely, they decompose $\boldsymbol{A}^{-1} = \bar{\mu}\boldsymbol{I}_n + \Delta_\mu$, where $\|\Delta_\mu\|_2 \ll \bar{\mu}$. This essentially shoves all the dependencies into Δ_μ. While the $\bar{\mu}$ portion reduces to the inner product calculation discussed above, they must use Cauchy-Schwarz to handle the dependent Δ_μ portion. This leads to inevitable looseness in the regimes for Theorem 2, as Cauchy-Schwarz is a *worst-case* bound. Interestingly, their Cauchy-Schwarz bound leverages the sparsity of the label vectors to gain a factor of $\sqrt{k}$, but the bound is still loose by a factor of $\sqrt{n}$.

We fully tighten the analysis for multiclass classification by directly analyzing the bilinear forms and fully exploiting their sparsity. To that end, we prove a variant of Hanson-Wright for *sparse bilinear forms*, which is quite similar to recent results about sparse bilinear and quadratic forms [30], [31] (recall that the original Hanson-Wright inequality only applies to *quadratic forms* that need not have sparsity). To actually apply Hanson-Wright, we carefully isolate the dependent portions using the Woodbury inversion formula, which generalizes the Sherman-Morrison formula for arbitrary rank updates.

Although our variant of Hanson-Wright is quite similar to Theorem 1 in [31], the assumptions are actually incomparable. The main proof techniques are heavily inspired by the style of analysis of [28], [30], [31]. To encode sparsity, we introduce the notation $\boldsymbol{v} \circ \boldsymbol{u} \in \mathbb{R}^n$ to denote the elementwise product of $\boldsymbol{v}, \boldsymbol{u} \in \mathbb{R}^n$. Then setting $\boldsymbol{v}$ to be a binary vector in $\{0,1\}^n$ allows us to explicitly encode sparsity.

Theorem 4 (Sparse Hanson-Wright for bilinear forms): Let $\boldsymbol{x} = (X_1, \dots, X_n) \in \mathbb{R}^n$ and $\boldsymbol{y} \in (Y_1, \dots, Y_n) \in \mathbb{R}^n$ be random vectors such that the pairs (X_i, Y_i) are independent pairs of (possibly correlated) centered random variables with subgaussian norm at most K. Suppose $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_n) \in \{0,1\}^n$ is an i.i.d. Bernoulli vector with bias π. Assume that $\boldsymbol{\gamma}$ is independent of $\boldsymbol{y}$, γ_j is independent of X_i for $i \neq j$, and finally *conditioned on* $\gamma_j = 1$, X_j has subgaussian norm at most K. Then there exists an absolute constant $c > 0$ such that for all $\boldsymbol{M} \in \mathbb{R}^{n \times n}$ and $\epsilon \geq 0$ we have

$$\begin{aligned}&\mathbb{P}\left[|\boldsymbol{x}^\top \boldsymbol{M}(\boldsymbol{y} \circ \gamma) - \mathbb{E}[\boldsymbol{x}^\top \boldsymbol{M}(\boldsymbol{y} \circ \gamma)]| > \epsilon\right] \\ &\quad \leq 2\exp\left(-c\min\left\{\frac{\epsilon^2}{K^4 \pi \|\boldsymbol{M}\|_F^2}, \frac{\epsilon}{K^2\|\boldsymbol{M}\|_2}\right\}\right).\end{aligned} \tag{28}$$

Note that the sparsity level π improves the concentration radius ϵ (smaller ϵ is better), which can be interpreted as the high probability bound. Since $\|\boldsymbol{M}\|_F^2 \leq n\|\boldsymbol{M}\|_2^2$, and $\pi = O(1/k)$ in our setting, we obtain a concentration radius ϵ which scales like $\sqrt{n/k}$ rather than $\sqrt{n}$ (obtained via Hanson-Wright without sparsity) or $n/\sqrt{k}$ (obtained via Cauchy-Schwarz with sparsity). This gain is crucial to tightly analyzing the survival and contamination terms.

B. Completing the proof sketch

Theorem 4 and the above insights about sparsity and independence allow us to prove (see Appendix) the following bounds on the relative survival and contamination terms.

Proposition 5 (Upper bound on relative survival): Suppose we are in the bi-level model in the regime where regression fails, i.e. $q + r > 1$. With probability at least $1 - O(1/n)$,

$$\lambda_F \left|\widehat{\boldsymbol{h}}_{\alpha,\beta}[\alpha]\right| \leq O(n^{1-q-r-\min\{t, \frac{1}{2}\}})\sqrt{\log k}.$$

Translating the parameters in Proposition 5, we see that the relative survival is diminished by a factor $1/k$ as long as $k = o(\sqrt{n})$, and a factor $1/\sqrt{n}$ for $k = \Omega(\sqrt{n})$. This roughly matches the expected behavior from the heuristic calculation. The looseness in the $k = \Omega(\sqrt{n})$ regime does not affect our final result. We now state our lower bound on contamination.

Proposition 6 (Lower bound on contamination): Suppose we are in the bi-level model regime where regression fails, i.e. $q + r > 1$. Then with probability at least $1 - O(1/n)$, the contamination satisfies

$$\mathsf{CN}_{\alpha,\beta} \geq \underbrace{\Omega(n^{1-q-r+\frac{r-t-1}{2}})}_{\mathsf{CN}_{\alpha,\beta,F}} + \underbrace{\Omega(n^{\frac{1-t-p}{2}})}_{\mathsf{CN}_{\alpha,\beta,U}}. \tag{29}$$

The first term $n^{1-q-r+\frac{r-t-1}{2}}$ arises from the contamination from favored[6] features $\mathsf{CN}_{\alpha,\beta,F}$. In the regime $t < \frac{1}{2}$, comparing the relative survival to this favored contamination yields a ratio of $n^{\frac{1-r-t}{2}}$. This decays polynomially exactly when $t > 1 - r$. On the other hand, the second term $n^{\frac{1-t-p}{2}}$ arises from the contamination from unfavored features $\mathsf{CN}_{\alpha,\beta,U}$. In the regime $t < \frac{1}{2}$, comparing the relative survival to this unfavored contamination yields a ratio of $n^{\frac{p+1-2(q+r)-t}{2}}$, which decays polynomially exactly when $t > p + 1 - 2(q + r)$. For $t \geq \frac{1}{2}$, we get a ratio of $n^{\frac{t-r}{2}}$, which shrinks because $t < r$. This explains the regimes for misclassification in Theorem 1.

V. Discussion

In this paper we present a weak-converse style impossibility result for *min-norm interpolative* multiclass classification using the overparameterized bi-level model, matching what was conjectured in [1]. This suggests that it might be possible to get a more general information-theoretic converse that limits the performance of any learning scheme that does not fully know the underlying covariance Σ of input features. Since entropic quantities in jointly normal contexts parallel the behavior of second-order quadratic forms involving correlations and covariances [32], the style of analysis here might be further developed to unlock such results.

More speculatively, we suspect that the classification problem here can be connected to an ultrawideband variant [33]–[35] of covert communication [36]–[41] in which the covert transmitter/receiver are connected by a wide fading channel and have only coordinated with a prior agreement to send pilot transmissions of the k distinct codewords before sending any new messages. The regime where regression doesn't work in learning is closely related to when the communication will be covert. Classification working using a learned codebook is required for the receiver to actually decode covert transmissions. Making such a connection precise might help bring new tools and perspectives to bear on such problems.

---

[6] Strictly speaking, this lower bound comes from the $s - k$ favored but not label defining features, but the distinction is asymptotically negligible.

REFERENCES

[1] V. Subramanian, R. Arya, and A. Sahai, "Generalization for multiclass classification with overparameterized linear models," in *Advances in Neural Information Processing Systems*, A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, Eds., 2022. [Online]. Available: https://openreview.net/forum?id=ikWvMRVQBWW

[2] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," *arXiv preprint arXiv:1611.03530*, 2016.

[3] ——, "Understanding deep learning (still) requires rethinking generalization," *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, 2021.

[4] P. L. Bartlett, A. Montanari, and A. Rakhlin, "Deep learning: a statistical viewpoint," *Acta numerica*, vol. 30, pp. 87–201, 2021.

[5] M. Belkin, "Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation," *Acta Numerica*, vol. 30, pp. 203–248, 2021.

[6] Y. Dar, V. Muthukumar, and R. G. Baraniuk, "A farewell to the bias-variance tradeoff? an overview of the theory of overparameterized machine learning," *arXiv preprint arXiv:2109.02355*, 2021.

[7] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani, "Surprises in high-dimensional ridgeless least squares interpolation," *The Annals of Statistics*, vol. 50, no. 2, pp. 949–986, 2022.

[8] S. Mei and A. Montanari, "The generalization error of random features regression: Precise asymptotics and the double descent curve," *Communications on Pure and Applied Mathematics*, vol. 75, no. 4, pp. 667–766, 2022.

[9] P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler, "Benign overfitting in linear regression," *Proceedings of the National Academy of Sciences*, vol. 117, no. 48, pp. 30 063–30 070, 2020.

[10] M. Belkin, D. Hsu, and J. Xu, "Two models of double descent for weak features," *SIAM Journal on Mathematics of Data Science*, vol. 2, no. 4, pp. 1167–1180, 2020.

[11] V. Muthukumar, K. Vodrahalli, V. Subramanian, and A. Sahai, "Harmless interpolation of noisy data in regression," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 67–83, 2020.

[12] D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro, "The implicit bias of gradient descent on separable data," *Journal of Machine Learning Research*, vol. 19, no. 1, pp. 2822–2878, 2018.

[13] Z. Ji and M. Telgarsky, "The implicit bias of gradient descent on nonseparable data," in *Conference on Learning Theory*, 2019, pp. 1772–1798.

[14] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of inverse problems*. Springer Science & Business Media, 1996, vol. 375.

[15] S. Gunasekar, J. Lee, D. Soudry, and N. Srebro, "Characterizing implicit bias in terms of optimization geometry," in *International Conference on Machine Learning*, 2018, pp. 1832–1841.

[16] V. Muthukumar, A. Narang, V. Subramanian, M. Belkin, D. J. Hsu, and A. Sahai, "Classification vs regression in overparameterized regimes: Does the loss function matter?" *Journal of Machine Learning Research*, vol. 22, pp. 222:1–222:69, 2021.

[17] N. S. Chatterji and P. M. Long, "Finite-sample analysis of interpolating linear classifiers in the overparameterized regime," *Journal of Machine Learning Research*, vol. 22, no. 129, pp. 1–30, 2021.

[18] K. Wang and C. Thrampoulidis, "Benign overfitting in binary classification of Gaussian mixtures," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 4030–4034.

[19] W. Wang and J. Fan, "Asymptotics of empirical eigenstructure for high dimensional spiked covariance," *Annals of statistics*, vol. 45, no. 3, p. 1342, 2017.

[20] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020.

[21] G. Wang, K. Donhauser, and F. Yang, "Tight bounds for minimum ℓ_1-norm interpolation of noisy data," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022, pp. 10 572–10 602.

[22] K. Wang, V. Muthukumar, and C. Thrampoulidis, "Benign Overfitting in Multiclass Classification: All Roads Lead to Interpolation," *arXiv e-prints*, p. arXiv:2106.10865, Jun. 2021.

[23] H. Hu and Y. M. Lu, "Universality laws for high-dimensional learning with random features," *IEEE Transactions on Information Theory*, 2022.

[24] Y. M. Lu and H.-T. Yau, "An equivalence principle for the spectrum of random inner-product kernel matrices," *arXiv preprint arXiv:2205.06308*, 2022.

[25] S. Goldt, B. Loureiro, G. Reeves, F. Krzakala, M. Mézard, and L. Zdeborová, "The Gaussian equivalence of generative models for learning with shallow neural networks," in *Mathematical and Scientific Machine Learning*. PMLR, 2022, pp. 426–471.

[26] T. Misiakiewicz, "Spectrum of inner-product kernel matrices in the polynomial regime and multiple descent phenomenon in kernel ridge regression," *arXiv preprint arXiv:2204.10425*, 2022.

[27] A. D. McRae, S. Karnik, M. Davenport, and V. K. Muthukumar, "Harmless interpolation in regression and classification with structured features," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022, pp. 5853–5875.

[28] M. Rudelson and R. Vershynin, "Hanson-Wright inequality and sub-Gaussian concentration," *Electronic Communications in Probability*, vol. 18, pp. 1–9, 2013.

[29] R. Vershynin, *High-dimensional probability: An introduction with applications in data science*. Cambridge university press, 2018, vol. 47.

[30] S. Zhou, "Sparse Hanson–Wright inequalities for subgaussian quadratic forms," *Bernoulli*, vol. 25, no. 3, pp. 1603–1639, 2019.

[31] S. Park, X. Wang, and J. Lim, "Sparse Hanson-Wright inequality for a Bilinear Form of Sub-Gaussian variables," *arXiv preprint arXiv:2209.05685*, 2022.

[32] A. Makur and L. Zheng, "Polynomial spectral decomposition of conditional expectation operators," in *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2016, pp. 633–640.

[33] M. Médard and R. G. Gallager, "Bandwidth scaling for fading multipath channels," *IEEE Transactions on Information Theory*, vol. 48, no. 4, pp. 840–852, 2002.

[34] D. Porrat, N. David, and S. Nacu, "Channel uncertainty in ultra-wideband communication systems," *IEEE Transactions on Information Theory*, vol. 53, no. 1, pp. 194–208, 2006.

[35] L. Zheng and D. N. C. Tse, "Communication on the Grassmann manifold: A geometric approach to the noncoherent multiple-antenna channel," *IEEE transactions on Information Theory*, vol. 48, no. 2, pp. 359–383, 2002.

[36] B. A. Bash, D. Goeckel, and D. Towsley, "Limits of reliable communication with low probability of detection on AWGN channels," *IEEE journal on selected areas in communications*, vol. 31, no. 9, pp. 1921–1930, 2013.

[37] B. A. Bash, D. Goeckel, D. Towsley, and S. Guha, "Hiding information in noise: Fundamental limits of covert wireless communication," *IEEE Communications Magazine*, vol. 53, no. 12, pp. 26–31, 2015.

[38] D. Goeckel, B. Bash, S. Guha, and D. Towsley, "Covert communications when the warden does not know the background noise power," *IEEE Communications Letters*, vol. 20, no. 2, pp. 236–239, 2015.

[39] L. Wang, G. W. Wornell, and L. Zheng, "Fundamental limits of communication with low probability of detection," *IEEE Transactions on Information Theory*, vol. 62, no. 6, pp. 3493–3503, 2016.

[40] Q. E. Zhang, M. R. Bloch, M. Bakshi, and S. Jaggi, "Undetectable radios: Covert communication under spectral mask constraints," in *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2019, pp. 992–996.

[41] Q. Zhang, M. Bakshi, and S. Jaggi, "Covert communication over adversarially jammed channels," *IEEE Transactions on Information Theory*, vol. 67, no. 9, pp. 6096–6121, 2021.