A Minimax Lower Bound for Low-Rank Matrix-Variate Logistic Regression

Batoul Taki, Mohsen Ghassemi, Anand D. Sarwate, and Waheed U. Bajwa
Department of Electrical and Computer Engineering, Rutgers University–New Brunswick, Piscataway, NJ

E-mail:{batoul.taki, mohsen.ghassemi, anand.sarwate, waheed.bajwa}@rutgers.edu

Abstract—**This paper considers the problem of matrix-variate logistic regression. This paper derives the fundamental error threshold on estimating low-rank coefficient matrices in the logistic regression problem by deriving a lower bound on the minimax risk. The bound depends explicitly on the dimension and distribution of the covariates, the rank and energy of the coefficient matrix, and the number of samples. The resulting bound is proportional to the intrinsic degrees of freedom in the problem, which suggests the sample complexity of the low-rank matrix logistic regression problem can be lower than that for vectorized logistic regression. The proof techniques utilized in this work also set the stage for development of minimax lower bounds for tensor-variate logistic regression problems.**

Index Terms—**logistic regression, low-rank matrix, minimax risk, singular value decomposition.**

I. INTRODUCTION

Logistic Regression (LR) is a statistical model used commonly in machine learning for classification problems [1]. In the simplest terms, LR seeks to model the conditional probability that a categorical response variable y_i takes on one of two possible values, $y_i \in \{0, 1\}$, which represents one of two possible events taking place (such as success/failure and detection/no detection). In regression analysis, the aim is to accurately estimate the model class parameters through a set of training data points $\{\mathbf{x}_i, y_i\}_{i=1}^n$, where $\mathbf{x}_i$ is the ith covariate sample vector. The conventional LR model is as follows:

$$\mathbb{P}_{y|\mathbf{x}}(y_i = 1|\mathbf{x}_i) = \frac{1}{1 + \exp^{-(\mathbf{b}^T \mathbf{x}_i + z)}}, \qquad (1)$$

where $y_i \in \{0, 1\}$ is the binary response, $\mathbf{x}_i \in \mathbb{R}^m$ is the known covariate vector, $\mathbf{b} \in \mathbb{R}^m$ is the deterministic but unknown coefficient vector, and z is the intercept where $\mathbb{E}[z] = 0$.

In many practical applications, covariates naturally take the form of two-dimensional arrays. Common examples include images, biological data such as electroencephalography (EEG), fiber bundle imaging [2] and spatio-temporal data [3]. Classical machine learning techniques vectorize such matrix covariates and estimate a coefficient vector. This leads to computational inefficiency and the destruction of the underlying spatial structure of the coefficient matrix, which often carries

valuable information [4], [5]. To address these issues, one can model the matrix LR problem as

$$\mathbb{P}_{y|\mathbf{x}}(y_i = 1|\mathbf{X}_i) = \frac{1}{1 + \exp^{-(\langle \mathbf{B}, \mathbf{X}_i \rangle + z)}}, \qquad (2)$$

where $\mathbf{X}_i \in \mathbb{R}^{m_1 \times m_2}$ is the known covariate matrix, $\mathbf{B} \in \mathbb{R}^{m_1 \times m_2}$ is the coefficient matrix, and z is the intercept. The model in (2) preserves the matrix structure of, and the valuable information in, the covariate data. In matrix-variate logistic regression analysis, the goal is to find an estimate of the coefficient $\mathbf{B}$.

In this paper, we focus on the high-dimensional setting where $n \ll m_1 m_2$ and derive a lower bound on the minimax risk of the matrix LR problem under the assumption that $\mathbf{B}$ has rank $r \ll \min\{m_1, m_2\}$. We note that the low-rank assumption has been used ubiquitously in regression analysis in former works [6], [7]. This low-rank structure may arise from the presence of latent variables, such that the model's intrinsic degrees of freedom are smaller than its extrinsic dimensionality. This allows us to represent the data in a lower-dimensional space and potentially reduce the sample complexity of estimating the parameters. Minimax lower bounds are useful in understanding the fundamental error thresholds of the problem under study, and in assessing the performance of the existing algorithms. They also provide beneficial insights to the parameters on which an achievable minimax risk might depend.

Whilst a number of works in the literature derive minimax lower bounds for higher-order linear regression problems [7], [8], as well as vector LR problems [6], [9], [10], to the best of our knowledge, little work has been done on this topic in matrix-variate LR problems. Regarding the existing literature, there are several works that study the matrix LR problem. Many works extend the matrix LR problem of (2) by proposing regularized matrix LR models in order to obtain rank-optimized or sparse estimates of the coefficient matrix in the model in (2) [5], [11]. Recently, works have introduced regularized matrix LR models for inference on image data [12]. Much of the literature develops efficient algorithms and provides empirical results on their performance [5], [11], [12] while some provide theoretical guarantees for the proposed algorithms [11]. By contrast, Baldin and Berthet [13] model the problem of graph regression as a matrix LR problem, where $\mathbf{B} \in \mathbb{R}^{m \times m}$. The matrix $\mathbf{B}$ is assumed to be block-

sparse, low-rank, and square. Under these assumptions (which are restrictive and not the most general), the authors derive minimax lower bounds on the estimation error $\left\|\widehat{\mathbf{B}} - \mathbf{B}\right\|_F^2$.

The three prior works [4], [11], [12] implement algorithms for matrix logistic regression but do not prove sample complexity bounds (upper or lower). In this paper, we derive a minimax lower bound on the error of a low-rank LR model which gives a bound on the number of samples necessary for estimating $\mathbf{B}$. Contrary to prior works, we impose minimal assumptions on $\mathbf{B}$, making our results generalizable to a larger class of matrices.

Our Minimax lower bound relies on the distribution of the covariates and the energy of the regression matrix. Following previous works that derive minimax lower bounds in the dictionary learning setting [14], [15], we use the standard strategy of lower bounding the minimax risk in parametric estimation problems by the maximum error probability in a multiple hypothesis test problem and using Fano's inequality [16]. We derive a lower bound that is proportional to the degrees of freedom in the rank-r matrix LR problem, and we reduce the sample complexity from $\mathcal{O}(m_1 m_2)$ in the vector setting to $\mathcal{O}(r(m_1 + m_2))$. This result is intuitively pleasing as it coincides with the number of free parameters in the model. We also show that our methods are easily extendable to the tensor case, i.e., $\underline{\mathbf{B}}$ is a coefficient tensor with some known properties.

II. MODEL AND PROBLEM FORMULATION

We use the following notation convention throughout the paper: x, $\mathbf{x}$, $\mathbf{X}$ and $\underline{\mathbf{X}}$ denote scalars, vectors, matrices and tensors, respectively. We denote $\mathbf{I}_m$ as the $m \times m$ identity matrix. Also, $\|\mathbf{x}\|_0$, $\|\mathbf{x}\|_2$ and $\|\mathbf{X}\|_F$ denote the ℓ_0, ℓ_2 and Frobenius norms, respectively. Given a matrix $\mathbf{X}$, $\mathbf{x}_j$ is the jth column of $\mathbf{X}$. If $\underline{\mathbf{X}}$ is a third-order tensor, then by fixing indices in the first and second modes, the matrix $\underline{\mathbf{X}}_{:,:,j}$ is the jth frontal slice of $\underline{\mathbf{X}}$. If $a \in \mathbb{R}$, then $\lfloor a \rfloor$ is the greatest integer less than a. We denote $\mathbf{A} \otimes \mathbf{C}$ as the Kronecker product of matrices $\mathbf{A}$ and $\mathbf{C}$. We define the function $\text{vec}(\mathbf{X})$ as the column-wise vectorization of matrix $\mathbf{X}$. Finally, $[R]$ is shorthand for $\{1, \ldots, R\}$.

Consider the matrix LR model in (2), in which the Bernoulli response $y_i \in \{0, 1\}$ is the ith response variable of n independent and identically distributed (i.i.d) observations. The covariate matrix $\mathbf{X}_i \in \mathbb{R}^{m_1 \times m_2}$ has independent, normally distributed, zero-mean and σ^2 variance entries. The response y_i is generated according to (2) from $\mathbf{X}_i$ and a fixed coefficient matrix $\mathbf{B} \in \mathbb{R}^{m_1 \times m_2}$ with $\|\mathbf{B}\|_F^2$ upper-bounded by a constant.

We consider the case where $\mathbf{B}$ is a rank-r matrix. Specifically, the rank-r singular value decomposition of $\mathbf{B}$ is

$$\mathbf{B} = \mathbf{B}_1 \mathbf{G} \mathbf{B}_2^T, \tag{3}$$

where $\mathbf{B}_1 \in \mathbb{R}^{m_1 \times r}$ and $\mathbf{B}_2 \in \mathbb{R}^{m_2 \times r}$ are (tall) singular vector matrices with orthonormal columns, and $\mathbf{G} =$

$\text{diag}(\lambda_1, \cdots, \lambda_r) \in \mathbb{R}^{r \times r}$ is the matrix of singular values with $\lambda_i > 0, \forall i \in [r]$. Under this low-rank structure of $\mathbf{B}$, we have

$$\mathbb{P}_{y|\mathbf{X}}(y_i = 1 | \mathbf{X}_i) = \frac{1}{1 + \exp^{-(\langle \mathbf{B}_1 \mathbf{G} \mathbf{B}_2^T, \mathbf{X}_i \rangle + z)}}. \tag{4}$$

We use Kronecker product properties [17] to express $\mathbf{b} = \text{vec}(\mathbf{B})$ as $\mathbf{b} = (\mathbf{B}_2 \otimes \mathbf{B}_1) \text{vec}(\mathbf{G})$.

The goal in LR is to find an estimate $\widehat{\mathbf{B}}$ of $\mathbf{B}$ using the training data $\{\mathbf{X}_i, y_i\}_{i=1}^n$. We assume that the true coefficient matrix $\mathbf{B}$ belongs to the set

$$\mathcal{B}_d(\mathbf{0}) \triangleq \{\mathbf{B}' \in \mathcal{P}_r : \rho(\mathbf{B}', \mathbf{0}) < d\}, \tag{5}$$

the open ball of radius d with distance metric $\rho = \|\cdot\|_F$, which resides in a given the parameter space,

$$\mathcal{P}_r \triangleq \Bigg\{ \mathbf{B}' \in \mathbb{R}^{m_1 \times m_2} : \text{rank}(\mathbf{B}') = r, \ \mathbf{B}' = \mathbf{B}_1' \mathbf{G}' \mathbf{B}_2', $$
$$\left\|\mathbf{b}_{1,j}'\right\|_2 = \left\|\mathbf{b}_{2,j}'\right\|_2 = 1, \mathbf{b}_{k,j}' \perp \mathbf{b}_{k,j'}', \forall j, j' \in [r],$$
$$j \neq j', k \in \{1, 2\} \Bigg\} \cup \{\mathbf{0}\}. \tag{6}$$

Thus $\mathcal{B}_d(\mathbf{0}) \subset \mathcal{P}_r$ and $\mathbf{B}$ has energy bounded by $\|\mathbf{B}\|_F^2 < d^2$.

Our objective is to find a lower bound on the minimax risk for estimating coefficient matrix $\mathbf{B}$. We define the non-decreasing function $\phi = \|\cdot\|_F^2 \triangleq \mathbb{R}_+ \rightarrow \mathbb{R}_+$ with $\phi(0) = 0$. The minimax risk is thus defined as the worst-case mean squared error (MSE) for the best estimator, i.e.,

$$\varepsilon^* = \inf_{\widehat{\mathbf{B}}} \sup_{\mathbf{B} \in \mathcal{B}_d(\mathbf{0})} \mathbb{E}_{\mathbf{y}, \underline{\mathbf{X}}^c} \left\{ \left\|\widehat{\mathbf{B}} - \mathbf{B}\right\|_F^2 \right\}, \tag{7}$$

where $\underline{\mathbf{X}}^c$ is the covariate tensor of n samples. We point out that minimax risk in (7) is an inherent property of the LR problem and holds for all possible estimators.

III. MINIMAX LOWER BOUND FOR LOW-RANK MATRIX LOGISTIC REGRESSION

The minimax lower bounds in the low-rank matrix LR setting that are based on Fano's inequality will depend explicitly on the dimensions of the covariate matrices $\mathbf{X}_i$ and their distribution, the rank, the upper bound on the energy of the coefficient matrix $\mathbf{B}$, the number of samples, n, and the construction of the multiple hypothesis test set.

The novelty of our work is that we explicitly leverage the rank-r structure of the coefficient matrix $\mathbf{B}$, in (3), which leads to the construction of each hypothesis that is structurally different in our setting compared to vector LR [6], [9], [10]. Furthermore, existing low-rank matrix packings, such as those in Negahban and Wainwright [18] are not useful in the LR setting of this paper, since the logistic function in our model makes part of our analysis non-trivial and fundamentally different to such works. In this work, we create a set $\mathcal{B}_L$ from three constructed sets (for $\mathbf{B}_1$, $\mathbf{B}_2$ and $\mathbf{G}$), and derive conditions under which all sets can exist. These conditions ensure the existence of a hypothesis set, $\mathcal{B}_L$, with elements that have a rank-r matrix structure and additional essential properties noted below. We also show that our chosen

constructed hypothesis set and analysis are more suitable for our matrix-variate model because they are easily generalizable to the tensor setting for LR.

A. Main Result

We derive lower bounds on ε^* using an argument based on Fano's Inequality [19]. To do so, we first relate the minimax risk in (7) to a multi-way hypothesis testing problem between an exponentially large family, with respect to the dimensions of the matrices, of L distinct coefficient matrices with energy upper bounded by d^2, i.e., matrices residing inside the openball $\mathcal{B}_d(\mathbf{0})$ of fixed radius:

$$\mathcal{B}_L = \{\mathbf{B}_l : l \in [L]\} \subset \mathcal{B}_d(\mathbf{0}), \tag{8}$$

where the correct hypothesis $\mathbf{B}_l$ is generated uniformly at random from the set $\mathcal{B}_L$.

More specifically, suppose there exists an estimator with worst-case MSE that matches ε^*. This estimator can be used to solve a multiple hypothesis testing problem using a minimum distance decoder. The minimax risk can then be lower bounded by the probability of error in such a multiple hypothesis test. Our main challenge is to further lower bound this probability of error in order to derive lower bounds on the minimax risk.

We now state our main result on the minimax risk of the low-rank matrix coefficient estimation problem in LR.

Theorem 1. *Consider the rank-r matrix LR problem in* (4) *with n i.i.d observations, $\{\mathbf{X}_i, y_i\}_{i=1}^{n}$ where the true coefficient matrix $\|\mathbf{B}\|_F^2 < d^2$. Then, for covariate $\text{vec}(\mathbf{X}_i) \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_m)$, the minimax risk is lower bounded by*

$$\varepsilon^* \geq \frac{\left(\left[c_2\big(c_1 r(m_1 + m_2 - 2) + c_1(r-1)\big) - c_3\right] - 1\right)}{8 n \sigma \sqrt{\frac{2}{\pi}}}, \tag{9}$$

where $c_1 = \left(1 - \frac{1}{10}\right)^2$, $c_2 = \frac{\log_2(e)(\sqrt{2}-1)}{4\sqrt{2}}$ *and* $c_3 = \left(\frac{3(\sqrt{2}-1)}{\sqrt{8}}\right)\log_2\left(\frac{3}{2}\right)$.

We discuss the implications of (9) of Theorem 1 in Section V

B. Roadmap for Theorem 1

For notational convenience, we define the response vector of n samples as $\mathbf{y} = [y_1, \cdots, y_n]$, and the covariate tensor of n samples as $\underline{\mathbf{X}}^c \in R^{m_1 \times m_2 \times n}$, where the ith frontal slice, $\underline{\mathbf{X}}^c_{:,:,i}$, is the covariate matrix $\mathbf{X}_i$ of the ith sample.

For our analysis, we construct a hypothesis set of L distinct matrices, as defined in (8). However, each hypothesis $\mathbf{B}_l$ is a low-rank matrix following the decomposition in (3) and is composed of three separately constructed sets (for $\mathbf{B}_1$, $\mathbf{B}_2$, and $\mathbf{G}$). Now, consider $\mathbb{I}(\mathbf{y}; l)$, the mutual information between the observations $\mathbf{y}$ and random index l. The construction of the set $\mathcal{B}_L$ is used to provide upper and lower bounds on $\mathbb{I}(\mathbf{y}; l)$, from which we derive lower bounds on the minimax risk ε^*.

For finding a lower bound on $\mathbb{I}(\mathbf{y}; l)$, we first consider the exponentially large packing set $\mathcal{B}_L$, with respect to the

dimensions of $\mathbf{B}$, where any two distinct hypotheses $\mathbf{B}_l$ and $\mathbf{B}_{l'}$ are separated by a minimum distance, i.e.,

$$\|\mathbf{B}_l - \mathbf{B}_{l'}\|_F^2 \geq 8\delta \tag{10}$$

for some positive δ. To achieve our desired bounds on the minimax risk, we require the existence of an estimator producing estimate $\widehat{\mathbf{B}}$ and achieving the minimax bound $\varepsilon^* < \sqrt{\delta}$. We consider the minimum distance decoder

$$\widehat{l}(\mathbf{y}) \triangleq \argmin_{\mathbf{B}_{l'} \in \mathcal{B}_d(\mathbf{0})} \left\|\widehat{\mathbf{B}} - \mathbf{B}_{l'}\right\|_F^2, \tag{11}$$

which seeks to detect the correct coefficient matrix $\mathbf{B}_l$. We require the estimate $\widehat{\mathbf{B}}$ to satisfy $\left\|\widehat{\mathbf{B}} - \mathbf{B}_l\right\|_F^2 < \sqrt{2\delta}$, for the minimum distance decoder to detect $\mathbf{B}_l$ and for the probability of detection error to be $\mathbb{P}(\widehat{l}(\mathbf{y}) \neq l) = 0$. A detection error might occur when $\left\|\widehat{\mathbf{B}} - \mathbf{B}_l\right\|_F^2 \geq \sqrt{2\delta}$. The following bound relates the loss $\left\|\widehat{\mathbf{B}} - \mathbf{B}_l\right\|_F^2$ and the probability of error $\mathbb{P}(\widehat{l}(\mathbf{y}) \neq l)$:

$$\mathbb{P}(\widehat{l}(\mathbf{y}) \neq l) \leq \mathbb{P}\left(\left\|\widehat{\mathbf{B}} - \mathbf{B}_l\right\|_F^2 \geq \sqrt{2\delta}\right). \tag{12}$$

Next, (12) is used to obtain a lower bound on $\mathbb{I}(\mathbf{y}; l)$ using Fano's inequality stated below [16]:

$$\mathbb{I}(\mathbf{y}; l) \geq \left(1 - \mathbb{P}(\widehat{l}(\mathbf{y}) \neq l)\right)\log_2(L) - 1 \triangleq u_1. \tag{13}$$

Secondly, for finding an upper bound on $\mathbb{I}(\mathbf{y}; l)$, we recognize that the LR model produces response variables y_i that are Bernoulli random variables conditioned on $\mathbf{X}_i$. On the basis thereof, we evaluate the mutual information, conditioned on $\underline{\mathbf{X}}^c$, $\mathbb{I}(\mathbf{y}; l | \underline{\mathbf{X}}^c)$. Next, define $f_l(\mathbf{y} | \underline{\mathbf{X}}^c)$ as the conditional probability distribution of $\mathbf{y}$ given $\underline{\mathbf{X}}^c$ with coefficient matrix $\mathbf{B}_l$, and let D_{KL} be the relative entropy, between any two $f_l(\mathbf{y} | \underline{\mathbf{X}}^c)$ and $f_{l'}(\mathbf{y} | \underline{\mathbf{X}}^c)$, produced from two distinct $\mathbf{B}_l$, $\mathbf{B}_{l'}$. Due to the convexity of $-\log$, we upper bound D_{KL} as follows [14], [20]:

$$\mathbb{I}(\mathbf{y}; l | \underline{\mathbf{X}}^c) \leq \frac{1}{L^2} \sum_{l,l'} \mathbb{E}_{\underline{\mathbf{X}}^c} D_{KL}(f_l(\mathbf{y} | \mathbf{X}) \| f_{l'}(\mathbf{y} | \mathbf{X})) \triangleq u_2. \tag{14}$$

Lastly, we remark that from [14] we have $\mathbb{I}(\mathbf{y}; l) \leq \mathbb{I}(\mathbf{y}; l | \underline{\mathbf{X}}^c)$, and $\mathbb{I}(\mathbf{y}; l | \underline{\mathbf{X}}^c)$ is trivially lower bounded by (13). Thus, (13) and (14) give rise to upper and lower bounds on the conditional mutual information:

$$u_1 \leq \mathbb{I}(\mathbf{y}; l | \mathbf{X}) \leq u_2, \tag{15}$$

where u_1 and u_2 require evaluation, and from which we obtain a lower bound on the minimax risk.

IV. PROOF OF MAIN RESULT

The formal proof for Theorem 1 relies on three lemmas. Lemma 1 introduces the exponentially large, with respect to the dimensions, set of vectors from which we will later construct $\mathbf{G}$. The set is constructed as a subset of an $(r-1)$-dimensional hypercube with a required minimum distance

479

between any two distinct elements in the set. Lemma 1 bounds the probability that this minimum distance is violated.

Lemma 1. *Let $r > 0$ and $F \geq 2$. Consider the set of F vectors $\{\mathbf{s}_f \in \mathbb{R}^{r-1} : f \in [F]\}$, where each entry in vector $\mathbf{s}_f$ is an independent and identically distributed random variable taking values $\left\{-\frac{1}{\sqrt{r-1}}, +\frac{1}{\sqrt{r-1}}\right\}$ uniformly. The probability that there exists a distinct pair (f, f') such that $\|\mathbf{s}_f - \mathbf{s}_{f'}\|_0 < \frac{r-1}{20}$ is upper bounded as follows:*

$$\mathbb{P}(\exists (f, f') \in [F] \times [F], f \neq f' : \|\mathbf{s}_f - \mathbf{s}_{f'}\|_0 < \frac{r-1}{20})$$
$$\leq \exp\left[2\log(F) - \log(2) - \frac{1}{2}\left(1 - \frac{1}{10}\right)^2 (r-1)\right]. \quad (16)$$

The above is a direct result of a standard application of Mc-Diarmid's inequality [21]. Notice, however, that our technique differs from [22] because we are imposing minimum distance conditions in the Hamming metric, rather than conditions on the inner product, between vectors.

Similar to the above, the following corollary introduces a set of matrices from which we will construct $\mathbf{B}_1, \mathbf{B}_2$.

Corollary 1. *Let $m, r > 0$ and $P \geq 2$. Consider the set of P matrices $\{\mathbf{S}_p \in \mathbb{R}^{(m-1) \times r} : p \in [P]\}$, where each entry in matrix $\mathbf{S}_p$ is an independent and identically distributed random variable taking values $\left\{-\frac{1}{\sqrt{(m-1)r}}, +\frac{1}{\sqrt{(m-1)r}}\right\}$ uniformly. The probability that there exists a distinct pair (p, p') such that $\|\mathbf{S}_p - \mathbf{S}_{p'}\|_0 \leq \frac{(m-1)r}{20}$ is upper bounded as follows:*

$$\mathbb{P}(\exists (p, p') \in [P] \times [P], p \neq p' : \|\mathbf{S}_p - \mathbf{S}_{p'}\|_0 < \frac{(m-1)r}{20})$$
$$\leq \exp\left[2\log(P) - \log(2) - \frac{1}{2}\left(1 - \frac{1}{10}\right)^2 (m-1)r\right]. \quad (17)$$

From the results in Lemma 1 and Corollary 1, Lemma 2 derives conditions on L such that $\mathcal{B}_L$ with a certain set of properties exists, and constructs two sets of matrices: 1) $m_1 \times r$ and $m_2 \times r$ orthonormal matrices from the set of "generating" matrices defined in Corollary 1, and 2) $r \times r$ diagonal matrices from the set of "generating" vectors in Lemma 1. Each element $\mathbf{B}_l \in \mathcal{B}_L$ is constructed from the three sets defined above, according to the decomposition in (3). Next, we prove lower and upper bounds on the distance between any two distinct elements in $\mathcal{B}_L$. The lower bound determines the minimum packing distance between any two $\mathbf{B}_l, \mathbf{B}_{l'}$, whilst the upper bound is used to derive the results in Lemma 3.

Lemma 2. *There exists a collection of L matrices $B_L \triangleq \{\mathbf{B}_l : l \in [L]\} \subset \mathcal{B}_d(\mathbf{0})$ for some $d > 0$ of cardinality*

$$L = 2^{\lfloor \frac{\log_2(e)}{4}\left(\left(1 - \frac{1}{10}\right)^2 (r(m_1 + m_2 - 1) + \left(1 - \frac{1}{10}\right)^2 (r-1)\right) - \frac{3}{2}\log_2\left(\frac{3}{2}\right)\rfloor}$$
$$(18)$$

such that for any

$$\sqrt{\frac{8(r-1)}{r}} < \varepsilon \leq d\sqrt{\frac{r-1}{r}}, \quad (19)$$

we have

$$\frac{r\varepsilon^2}{r-1} < \|\mathbf{B}_l - \mathbf{B}_{l'}\|_F^2 \leq 4\frac{r\varepsilon^2}{r-1}. \quad (20)$$

Lemma 3 derives an upper bound on $\mathbb{I}(\mathbf{y}; l | \underline{\mathbf{X}}^c)$.

Lemma 3. *Consider the matrix LR problem given by the model in (2) such that $\mathbf{B} \in \mathcal{B}_d(\mathbf{0})$, for some $d > 0$, and consider the set $\mathcal{B}_L$ defined in Lemma 2. Consider n i.i.d observations y_i following a Bernoulli distribution when conditioned on $\mathbf{X}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$, then we have,*

$$\mathbb{I}(\mathbf{y}; l | \mathbf{X}) \leq n\frac{2}{r}\sqrt{\frac{2}{\pi}}\sigma\varepsilon. \quad (21)$$

The final lemma bounds the error probability $\mathbb{P}(\widehat{l}(\mathbf{y}) \neq l)$ under the conditions, mentioned in Section III-B, on ε^* and $\left\|\widehat{\mathbf{B}} - \mathbf{B}_l\right\|_F^2$ for the recovery of hypothesis $\mathbf{B}_l$ by the minimum distance decoder. The proof follows exactly that for Lemma 8 in [14].

Lemma 4. *Consider the minimum distance decoder in (11), for the L matrices constructed in Lemma 2. Consider the LR regression model in (2) with minimax risk ε^*, which is assumed to be upper bounded by $\varepsilon^* \leq \sqrt{\delta}$. The minimum distance decoder recovers the correct hypothesis if $\left\|\widehat{\mathbf{B}} - \mathbf{B}_l\right\|_F^2 < \sqrt{2\delta}$. The detection error of the minimum distance decoder is upper bounded by $\mathbb{P}(\widehat{l}(\mathbf{y}) \neq l) \leq \frac{1}{\sqrt{2}}$.*

Proof of Theorem 1: Fix r, let $\varepsilon > 0$ satisfy (19), and define $\varepsilon^2 \triangleq \frac{8\delta(r-1)}{r}$. Suppose there exists an estimator $\widehat{\mathbf{B}}$ which guarantees a risk $\varepsilon^* \leq \sqrt{\delta} = \sqrt{\frac{r}{8(r-1)}}\varepsilon$. By Lemma 2, there exists a packing set B_L containing L distinct rank-r matrices, where L satisfies (18) and for any pair $\mathbf{B}_l, \mathbf{B}_{l'} \in B_L$, the distance $\|\mathbf{B}_l - \mathbf{B}_{l'}\|_F^2$ satisfies (10). By Lemma 3, the conditional mutual information $\mathbb{I}(\mathbf{y}; l | \underline{\mathbf{X}}^c)$ is upper bounded by (21), and $\mathbb{P}(\widehat{l}(\mathbf{y}) - \mathbf{B}_l)$ is upper bounded by Lemma 4. Replacing (21) and (13) in (15), we have:

$$\frac{\sqrt{2} - 1}{\sqrt{2}}\log_2 L - 1 \leq \mathbb{I}(\mathbf{y}; l | \mathbf{X}) \leq n\sigma\frac{2}{r}\sqrt{\frac{2}{\pi}}\varepsilon. \quad (22)$$

Lastly, if we set $\varepsilon^* = \sqrt{\frac{r}{8(r-1)}}\varepsilon$, then we achieve the result in (9). ∎

V. DISCUSSION AND CONCLUSION

In this paper we provided a minimax lower bound on the low-rank matrix-variate LR problem in the high-dimensional setting. We constructed a packing set of low-rank structured matrices with finite energy. Using the construction, we derived bounds on the conditional mutual information defined in our problem, in order to obtain a lower bound on the minimax risk. Compared to the vector case, such as in [6], the result in Theorem 1 shows a decrease in the lower bound from

480

$\mathcal{O}(m_1 m_2)$ to $\mathcal{O}(r(m_1 + m_2))$. The result also shows that the lower bound on the minimax risk is proportional to the intrinsic degrees of freedom in the coefficient matrix LR, (i.e., $r(m_1 + m_2 + 1)$), and decreases with increasing sample size n. This suggests that we can develop algorithms that take advantage of the low-rank structure of the coefficient matrices. Moreover, the result in (9) can be generalized from low-rank matrices to low-rank tensors. Imposing a rank-r decomposition on a coefficient tensor, $\underline{\mathbf{B}}$, holds the same conveniences as those discussed above of low-rank matrices. This low-rank structure on $\underline{\mathbf{B}}$ is a special case of the well-known Canonical Polyadic Decomposition or Parallel Factors (CANDECOMP/PARAFAC, or CP) [17], formally defined as,

$$\underline{\mathbf{B}} := \sum_{h=1}^{r} \lambda_h \mathbf{b}_1^{(h)} \circ \cdots \circ \mathbf{b}_K^{(h)}, \tag{23}$$

where $\mathbf{b}_k^{(h)} \in \mathbb{R}^{m_k}$ is a column vector along the kth mode of $\underline{\mathbf{B}}$, $\circ$ is the outer product, and $\lambda_h \mathbf{b}_1^{(h)} \circ \cdots \circ \mathbf{b}_K^{(h)}$, for any rank h, is a rank-1 tensor weighted by λ_h. The rank-r CP-decomposition expresses tensors as a sum of r rank-1 tensors. Equivalent to (23) is the following expression of a CP-structured tensor:

$$\underline{\mathbf{B}} = \underline{\mathbf{G}} \times_1 \mathbf{B}_1 \times_2 \cdots \times_K \mathbf{B}_K, \tag{24}$$

where the tensor $\underline{\mathbf{G}} \in \mathbb{R}^{\overbrace{r \times \cdots \times r}^{K\text{-times}}}$ is simply a higher-order analogue of the diagonal matrix $\mathbf{G}$ in (3) (where only the elements along the super-diagonal of $\underline{\mathbf{G}}$ are non-zero), and the mode-k factor matrices $\mathbf{B}_k \in \mathbb{R}^{m_k \times r}$, $\forall k \in [K]$ are rank-r matrices with orthonormal columns. Thus, the setup and construction proposed in this paper can be extended to the tensor case, where $\mathbf{B}$ is simply a special case of $\underline{\mathbf{B}}$, with $K = 2$ factor matrices.

VI. APPENDIX

Proof of Lemma 1: Consider the set of F vectors $\{\mathbf{s}_f \in \mathbb{R}^{r-1} : f \in [F]\}$, in (1). For indices $f, f' \in [F]$, $f \neq f'$, define the vector $\widetilde{\mathbf{s}} \triangleq \mathbf{s}_f \odot \mathbf{s}_{f'}$ as the point-wise product of $\mathbf{s}_l$ and $\mathbf{s}_{l'}$, and $\widetilde{s}_i$ as the ith entry of $\widetilde{\mathbf{s}}$, for $i \in [r-1]$. Define also the function

$$h(\widetilde{\mathbf{s}}) \triangleq (r-1) \sum_{i=1}^{r-1} \widetilde{\mathbf{s}}_i. \tag{25}$$

We use $\widetilde{\mathbf{s}} \approx \widetilde{\mathbf{s}}'$ to say that $\widetilde{s}_i = \widetilde{s}_i'$ for all entries $i \in [r-1]$, except one.

We require a minimum packing distance

$$\|\mathbf{s}_f - \mathbf{s}_{f'}\|_0 = \frac{1}{2} \left[(r-1) - (r-1) \sum_{i=1}^{r-1} \widetilde{s}_i \right] > \frac{r-1}{20}. \tag{26}$$

The probability that the requirement in (26) is not satisfied is defined as

$$\mathbb{P}\left((r-1) \sum_{i=1}^{r} \widetilde{\mathbf{s}}_i \geq (r-1)(1 - \frac{1}{10}) \right). \tag{27}$$

The function $h(\cdot)$ satisfies

$$\sup_{\widetilde{\mathbf{s}} \approx \widetilde{\mathbf{s}}'} |h(\widetilde{\mathbf{s}}) - h(\widetilde{\mathbf{s}}')| = |(r-1)\frac{1}{r-1} + (r-1)\frac{1}{r-1}| = 2. \tag{28}$$

According to McDiarmid's inequality in [21], for $\frac{(r-1)(9)}{10} > 0$, and since $\mathbb{E}_{\widetilde{\mathbf{s}}}[h(\widetilde{\mathbf{s}})] = 0$, we have,

$$\mathbb{P}\left((r-1) \sum_{i=1}^{r-1} \widetilde{\mathbf{s}}_i \geq (r-1)\frac{9}{10} \right)$$

$$\leq \exp\left[-\frac{1}{2} \left(1 - \frac{1}{10} \right)^2 (r-1) \right]. \tag{29}$$

The probability in (29) is for any two distinct pairs $(f, f') \in [F] \times [F]$. We take a union bound over all $\binom{L}{2}$ distinct pairs and upper bound the probability as:

$$\mathbb{P}(\exists (f, f') \in [F] \times [F], f \neq f' : (r-1) \sum_{i=1}^{r-1} \widetilde{s}_i \geq \frac{(r-1)9}{10})$$

$$\leq \frac{F^2}{2} \exp\left[-\frac{1}{2} \left(1 - \frac{1}{10} \right)^2 (r-1) \right]$$

$$= \exp\left[2 \log(F) - \log(2) - \frac{1}{2} \left(1 - \frac{1}{10} \right)^2 (r-1) \right]. \tag{30}$$

∎

Proof of Lemma 2: Fix the following arbitrary real orthogonal bases: $\mathbf{Q}$ of $\mathbb{R}^r$, the set of distinct r bases, $\{\mathbf{U}_{1,j}\}_{j=1}^r$ of $\mathbb{R}^{m_1}$, and the set of distinct r bases, $\{\mathbf{U}_{2,j}\}_{j=1}^r$ of $\mathbb{R}^{m_2}$.

Next, consider the following hypercubes or subsets thereof:
1) The set of F vectors $\{\mathbf{s}_f\}$ from Lemma 1:

$$\mathbf{s}_f \in \left\{ \frac{-1}{\sqrt{r-1}}, \frac{+1}{\sqrt{r-1}} \right\} \subset \mathbb{R}^{r-1}, \tag{31}$$

2) Two sets of P_1 and P_2 matrices, from Corollary 1:

$$\mathbf{S}_{p_1} \in \left\{ \frac{-1}{\sqrt{(m_1-1)r}}, \frac{+1}{\sqrt{(m_1-1)r}} \right\} \subset \mathbb{R}^{(m_1-1) \times r}, \tag{32}$$

and

$$\mathbf{S}_{p_2} \in \left\{ \frac{-1}{\sqrt{(m_2-1)r}}, \frac{+1}{\sqrt{(m_2-1)r}} \right\} \subset \mathbb{R}^{(m_2-1) \times r}, \tag{33}$$

respectively.

From Lemma 1 we have the following bounds on the probability that the minimum distance condition is violated for the set (31):

$$\mathbb{P}\left(\exists (f, f') \in [F] \times [F], f \neq f' : \|\mathbf{s}_f - \mathbf{s}_{f'}\|_0 < \frac{r-1}{20} \right)$$

$$\leq \exp\left[\log(\frac{F^2}{2}) - \frac{(r-1)}{2} \left(1 - \frac{2}{20} \right)^2 \right]. \tag{34}$$

481

Likewise, from Corollary 1, we have the following bounds on the probability that the minimum distance conditions are violated for the sets (32) and (33), respectively:

$$\mathbb{P}\Bigg(\exists (p_1, p_1') \in [P_1] \times [P_1], p_1 \neq p_1' : \left\| \mathbf{S}_{p_1} - \mathbf{S}_{p_1'} \right\|_0$$

$$< \frac{(m_1 - 1)r}{20}\Bigg)$$

$$\leq \exp\left[\log\left(\frac{P_1{}^2}{2}\right) - \frac{(m_1 - 1)r}{2}\left(1 - \frac{2}{20}\right)^2\right], \quad (35)$$

and

$$\mathbb{P}\Bigg(\exists (p_2, p_2') \in [P_2] \times [P_2], p_2 \neq p_2' : \left\| \mathbf{S}_{p_2} - \mathbf{S}_{p_2'} \right\|_0$$

$$< \frac{(m_2 - 1)r}{20}\Bigg)$$

$$\leq \exp\left[\log\left(\frac{P_2{}^2}{2}\right) - \frac{(m_2 - 1)r}{2}\left(1 - \frac{2}{20}\right)^2\right]. \quad (36)$$

We require the set of coefficient matrices $\mathcal{B}_L$ from the sets in (31), (32) and (33) to exist simultaneously. Hence, using a union bound on (34), (35) and (35), we can choose parameters to guarantee the existence of a construction. This is satisfied if the following conditions on the cardinalities F, P_1 and P_2 hold:

$$0 < F < \frac{\log_2(e)}{4}\left(1 - \frac{1}{10}\right)^2 (r - 1) - \frac{1}{2}\log_2\left(\frac{3}{2}\right), \quad (37)$$

$$0 < P_1 < 2^{\frac{\log_2(e)}{4}\left(1 - \frac{1}{10}\right)^2 (m_1 - 1)r - \frac{1}{2}\log_2\left(\frac{3}{2}\right)}, \quad (38)$$

and

$$0 < P_2 < 2^{\frac{\log_2(e)}{4}\left(1 - \frac{1}{10}\right)^2 (m_2 - 1)r - \frac{1}{2}\log_2\left(\frac{3}{2}\right)}. \quad (39)$$

Note that (37), (38) and (39) are sufficient conditions for the simultaneous existence of sets in (31), (32) and (33), such that the minimum distance condition between any two elements in each set is satisfied.

We proceed with the following steps in order to construct the final set $\mathcal{B}_L$ of coefficient matrices. Without loss of generality, we assume that the energy of any $\mathbf{B}_l$ is upper bounded by d^2. We will construct diagonal matrices $\mathbf{G}_f$, and matrices with orthonormal columns, namely $\mathbf{B}_{1,p_1}$ and $\mathbf{B}_{1,p_2}$, all of which will be used to construct each $\mathbf{B}_l \in \mathcal{B}_L$. In other words, due to our LR model, any matrix $\mathbf{B}_l$ will have a rank-r singular value decomposition. Thus, a bound on the matrix norm of $\mathbf{B}_l$ gives a bound on the norm of the matrix of singular values.

Firstly, we construct vectors $\mathbf{g}_f^{(1)} \in \mathbb{R}^r$ for $f \in [F]$, using $\mathbf{Q}$ and $\mathbf{s}_f$, as follows:

$$\mathbf{g}_f^{(1)} = \mathbf{Q}\begin{bmatrix} \sqrt{\frac{1}{r-1}} \\ \mathbf{s}_f \end{bmatrix}, \forall f \in [F]. \quad (40)$$

From (40), since $\|\mathbf{s}_f\|^2 = 1$ we have:

$$\left\|\mathbf{g}_f^{(1)}\right\|_2^2 = \left\|\mathbf{Q}\begin{bmatrix} \sqrt{\frac{1}{r-1}} \\ \mathbf{s}_f \end{bmatrix}\right\|_2^2 = \frac{r}{r-1}.$$

Similarly, we construct matrices $\mathbf{B}_{1,p_1}^{(1)} \in \mathbb{R}^{m_1 \times r}$, for $p_1 \in [P_1]$ and $\mathbf{B}_{2,p_2}^{(1)} \in \mathbb{R}^{m_2 \times r}$, for $p_2 \in [P_2]$, respectively. Define $\mathbf{b}_{1,p_1,j}^{(1)}$ as the jth column of $\mathbf{B}_{1,p_1}^{(1)}$, and $\mathbf{b}_{1,p_1,j}^{(2)}$ as the jth column of $\mathbf{B}_{2,p_2}^{(1)}$. Let the columns be constructed as follows:

$$\mathbf{b}_{1,p_1,j}^{(1)} = \mathbf{U}_{1,j}\begin{bmatrix} 1 \\ \mathbf{s}_{p_1,j} \end{bmatrix}, \forall p_1 \in [P_1], \quad (41)$$

and

$$\mathbf{b}_{2,p_2,j}^{(1)} = \mathbf{U}_{2,j}\begin{bmatrix} 1 \\ \mathbf{s}_{p_2,j} \end{bmatrix}, \forall p_2 \in [P_2]. \quad (42)$$

From (41) and (42), we have:

$$\left\|\mathbf{b}_{1,p_1,j}^{(1)}\right\|_2^2 = \left\|\mathbf{b}_{2,p_2,j}^{(1)}\right\|_2^2 = \left\|\begin{bmatrix} 1 \\ \mathbf{s}_{p_1,j} \end{bmatrix}\right\|_2^2 = \frac{r+1}{r}.$$

Secondly, we construct an r-sparse vector $\mathbf{g}_f^{(2)} \in \mathbb{R}^{r^2}$, element-wise, from $\mathbf{g}_f^{(1)}$. Define $g_{f,i}^{(2)}$ as the ith element of $\mathbf{g}_f^{(2)}$, where $i \in [r^2]$ and use the following construction:

$$g_{f,i}^{(2)} = \begin{cases} |g_{f,i}^{(1)}| & i = i' + r(i' - 1), i' = \{1, \ldots, r\} \\ 0 & \text{otherwise} \end{cases}, \quad (43)$$

and we note that

$$\left\|\mathbf{g}_f^{(2)}\right\|_2^2 = \left\|\mathbf{g}_f^{(1)}\right\|_2^2 = \frac{r}{r-1}. \quad (44)$$

We also construct matrices $\mathbf{B}_{1,p_1}^{(2)} \in \mathbb{R}^{m_1 \times r}$, for $p_1 \in [P_1]$ and $\mathbf{B}_{2,p_2}^{(2)} \in \mathbb{R}^{m_2 \times r}$, for $p_2 \in [P_2]$. We show the construction of $\mathbf{B}_{1,p_1}^{(2)}$ only: the construction of $\mathbf{B}_{2,p_2}^{(2)}$ follows the same procedure. Define $\mathbf{b}_{1,p_1,j}^{(2)} \in \mathbb{R}^{m_1}$ as the jth column of $\mathbf{B}_{1,p_1}^{(2)}$, for $j \in [r]$. We set

$$\mathbf{b}_{1,p_1,1}^{(2)} = \frac{\mathbf{b}_{1,p_1,1}^{(1)}}{\left\|\mathbf{b}_{1,p_1,1}^{(1)}\right\|_2}, \quad (45)$$

and define

$$\mathbf{a}_{j+1} \triangleq \mathbf{b}_{1,p_1,j+1}^{(1)} - \sum_{j'=1}^{j} \langle \mathbf{b}_{1,p_1,j+1}^{(1)}, \mathbf{b}_{1,p_1,j'}^{(2)} \rangle \mathbf{b}_{1,p_1,j'}^{(2)}, \quad (46)$$

and

$$\mathbf{b}_{1,p_1,j+1}^{(2)} \triangleq \frac{\mathbf{a}_{j+1}}{\|\mathbf{a}_{j+1}\|_2}, \quad (47)$$

for $j \in [r - 1]$.

The steps in (45), (46) and (47) constitute the well-known Gram-Schmidt Process. Thus, set of vectors $\mathbf{b}_{1,p_1,j}^{(2)}$, for $j \in [r - 1]$, $p_1 \in [P_1]$ are orthonormal, i.e, $\left\|\mathbf{b}_{1,p_1,j}^{(2)}\right\|_2^2 = 1$ and $\mathbf{b}_{1,p_1,j}^{(2)} \perp \mathbf{b}_{1,p_1,j'}^{(2)}$, for any two distinct $j, j' \in [r]$. Consequently, $\left(\mathbf{B}_{2,p_2}^{(2)}\right)^T \left(\mathbf{B}_{2,p_2}^{(2)}\right) = \mathbf{I}_r$.

Finally, we define the vector $\mathbf{g}_f$ and matrices $\mathbf{B}_{1,p_1}$ and $\mathbf{B}_{1,p_2}$ as:

$$\mathbf{g}_f = \frac{\varepsilon}{r}\mathbf{g}_f^{(2)}, \forall f \in [F], \quad (48)$$

482

and

$$\mathbf{B}_{1,p_1} = \mathbf{B}_{1,p_1}^{(2)}, \ \ \mathbf{B}_{1,p_2} = \mathbf{B}_{2,p_2}^{(2)}, \tag{49}$$

respectively, for some positive number ε.

We also define the diagonal matrix $\mathbf{G}_f \in \mathbb{R}^{r \times r}$, where $\text{vec}(\mathbf{G}_f) = \mathbf{g}_f$.

Now, by designating

$$\mathcal{L} \triangleq \{(f,p_1,p_2) : f \in [F], p_1 \in [P_1], p_2 \in [P_2]\}, \tag{50}$$

as the set of possible tuples (f,p_1,p_2), we have

$$L = |\mathcal{L}| \overset{(a)}{<} 2^{\left[\frac{\log_2(e)}{4}(c_1(r(m_1+m_2-2)+(r-1)))-(\frac{3}{2})\log_2(\frac{3}{2})\right]}, \tag{51}$$

where (a) follows from (37), (38) and (39). We define the set of coefficient matrices, $\mathcal{B}_L$ as,

$$\mathcal{B}_L \triangleq \left\{ \mathbf{B}_l = \mathbf{B}_{1,p_1}^T \mathbf{G}_f \mathbf{B}_{1,p_2}, l \in [L], f \in [F], \right. \tag{52}$$

$$\left. p_1 \in [P_1], p_2 \in [P_2] \right\}, \tag{53}$$

and we restrict ε such that

$$\sqrt{\frac{8(r-1)}{r}} < \varepsilon < d\sqrt{\frac{r-1}{r}}, \tag{54}$$

in order to guarantee $\|\mathbf{G}\|_2^2 < \frac{d^2}{r^2}$. We make the final note that, due to the Kronecker product, we can express $\text{vec}(\mathbf{B}_l)$ as:

$$\text{vec}(\mathbf{B}_l) = (\mathbf{B}_{1,p_1} \otimes \mathbf{B}_{1,p_2})\mathbf{g}_f. \tag{55}$$

We have the following remaining tasks at hand: 1) We must show that they energy of any $\mathbf{B}_l$ is less than d^2. 2) We must derive upper and lower bounds on the distance between any two distinct $\mathbf{B}_l, \mathbf{B}_{l'} \in B_L$ $\left(\|\mathbf{B}_l - \mathbf{B}_{l'}\|_F^2\right)$.

We begin by showing $\|\mathbf{B}_l\|_F^2 < d^2$:

$$\begin{aligned}
\|\mathbf{B}_l\|_F^2 &= \|(\mathbf{B}_{1,p_2} \otimes \mathbf{B}_{1,p_1})\mathbf{g}_f\|_2^2 \\
&\leq \|\mathbf{B}_{1,p_2} \otimes \mathbf{B}_{1,p_1}\|_F^2 \|\mathbf{g}_f\|_2^2 \\
&\overset{(b)}{=} \|\mathbf{B}_{1,p_1}\|_F^2 \|\mathbf{B}_{1,p_2}\|_F^2 \|\mathbf{g}_f\|_2^2 = \frac{r\varepsilon^2}{r-1} \overset{(c)}{<} d^2,
\end{aligned}$$

where (b) follows from the fact that the matrix norm of the Kronecker product is the product of the matrix norms, and (c) holds due to (54).

We proceed with deriving lower and upper bounds on $\|\mathbf{B}_l - \mathbf{B}_{l'}\|_F^2$, for any two distinct $\mathbf{B}_l, \mathbf{B}_{l'} \in B_L$. For finding lower bounds on $\|\mathbf{B}_l - \mathbf{B}_{l'}\|_F^2$, it can be shown that the closest pair $\mathbf{B}_l, \mathbf{B}_{l'}$ occurs for $p_1 = p_1'$, $p_2 = p_2'$ and $f \neq f'$. Thus we have the following:

$$\begin{aligned}
\|\mathbf{B}_l - \mathbf{B}_{l'}\|_F^2 &= \left\|(\mathbf{B}_{1,p_2} \otimes \mathbf{B}_{1,p_1})\mathbf{g}_f - (\mathbf{B}_{1,p_2'} \otimes \mathbf{B}_{1,p_1'})\mathbf{g}_{f'}\right\|_2^2 \\
&\geq \|(\mathbf{B}_{1,p_2} \otimes \mathbf{B}_{1,p_1})(\mathbf{g}_f - \mathbf{g}_{f'})\|_2^2 \\
&\overset{(d)}{=} \frac{\varepsilon^2}{r^2}\left\|\mathbf{g}_f^{(2)} - \mathbf{g}_{f'}^{(2)}\right\|_2^2 \\
&> \frac{r\varepsilon^2}{r-1}\frac{4}{20}
\end{aligned}$$

where (d) follows from the fact that the Kronecker product of orthogonal bases is an orthogonal basis.

Finally, for finding upper bounds on $\|\mathbf{B}_l - \mathbf{B}_{l'}\|_F^2$, we have:

$$\begin{aligned}
&\|\mathbf{B}_l - \mathbf{B}_{l'}\|_F^2 \\
&\overset{(e)}{\leq} \left[\|(\mathbf{B}_{1,p_2} \otimes \mathbf{B}_{1,p_1})\mathbf{g}_f\|_F + \|(\mathbf{B}_{1,p_2'} \otimes \mathbf{B}_{1,p_1'})\mathbf{g}_{f'}\|_F\right]^2 \\
&\leq \left[\|\mathbf{B}_{1,p_2} \otimes \mathbf{B}_{1,p_1}\|_F \|\mathbf{g}_f\|_2 + \|\mathbf{B}_{1,p_2'} \otimes \mathbf{B}_{1,p_1'}\|_F \|\mathbf{g}_{f'}\|_2\right]^2 \\
&= \left[2\|\mathbf{B}_{1,p_1}\|_F \|\mathbf{B}_{1,p_2}\|_F \|\mathbf{g}_f\|_2\right]^2 \tag{56} \\
&\leq 4\frac{r\varepsilon^2}{r-1} \tag{57}
\end{aligned}$$

where (e) follows from the triangle inequality. $\blacksquare$

Proof of Lemma 3: Consider the set $\mathcal{B}_L$ defined in Lemma 2, where the bounds in (56) and (57) hold. Consider the matrix LR model in (2). For n i.i.d samples, consider covariate matrices $\mathbf{X}_i \in \mathbb{R}^{m_1 \times m_2}, \forall i \in [n]$, where $\text{vec}(\mathbf{X}_i) \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{m_1 m_2})$. According to (2), observations y_i follow a Bernoulli distribution when conditioned on $\mathbf{X}_i$, $\forall i \in [n]$. Consider $\mathbf{y}$ and $\underline{\mathbf{X}}^c$ defined in Section II, and define $\mathbb{I}(\mathbf{y}; l|\underline{\mathbf{X}}^c)$ as the mutual information between observations $\mathbf{y}$ and index l conditioned on side-information $\underline{\mathbf{X}}^c$. From [20], [23], we have

$$\mathbb{I}(\mathbf{y}; l|\underline{\mathbf{X}}^c) \leq \frac{1}{L^2}\sum_{l,l'} \mathbb{E}_{\underline{\mathbf{X}}^c} D_{KL}(f_l(\mathbf{y}|\underline{\mathbf{X}}^c)||f_{l'}(\mathbf{y}|\underline{\mathbf{X}}^c), \tag{58}$$

where $D_{KL}(f_l(\mathbf{y}|\underline{\mathbf{X}}^c)||f_{l'}(\mathbf{y}|\underline{\mathbf{X}}^c)$ is the Kullback-Leibler (KL) divergence of probability distribution $f_l(\mathbf{y}|\underline{\mathbf{X}}^c)$ of $\mathbf{y}$ given $\underline{\mathbf{X}}^c$ for some $\mathbf{B}_l \in \mathcal{B}_L$. Denote $\sigma_l \triangleq \frac{1}{1+\exp^{-\langle\mathbf{B}_l,\mathbf{X}_i\rangle}}$, and $\sigma_{l'} \triangleq \frac{1}{1+\exp^{-\langle\mathbf{B}_{l'},\mathbf{X}_i\rangle}}$, We evaluate the KL divergence as follows:

$$\begin{aligned}
&D_{KL}(f_l(\mathbf{y}|\underline{\mathbf{X}}^c)||f_{l'}(\mathbf{y}|\underline{\mathbf{X}}^c)) \\
&= \sum_{i\in[n]} \sigma_l \log\left(\frac{\sigma_l}{\sigma_{l'}}\right) + (1-\sigma_l)\log\left(\frac{1-\sigma_l}{1-\sigma_{l'}}\right). \\
&= \sum_{i\in[n]} \sigma_l\langle\mathbf{B}_l,\mathbf{X}_i\rangle - \sigma_l\langle\mathbf{B}_{l'},\mathbf{X}_i\rangle - \langle\mathbf{B}_l,\mathbf{X}_i\rangle + \langle\mathbf{B}_{l'},\mathbf{X}_i\rangle \\
&\quad + \log(1+\exp^{-\langle\mathbf{B}_l,\mathbf{X}_i\rangle}) - \log(1+\exp^{-\langle\mathbf{B}_{l'},\mathbf{X}_i\rangle}). \tag{59}
\end{aligned}$$

Now, considering the distribution on covariates $\mathbf{X}_i, \forall i \in [n]$, we take the expectation of (59) with respect to the side-information $\underline{\mathbf{X}}^c$. We have $\mathbb{E}_{\underline{\mathbf{X}}^c}(\log(1+\exp^{-\langle\mathbf{B}_l,\mathbf{X}_i\rangle})) = \mathbb{E}_{\underline{\mathbf{X}}^c}(\log(1+\exp^{-\langle\mathbf{B}_{l'},\mathbf{X}_i\rangle}))$, and $\mathbb{E}_{\underline{\mathbf{X}}^c}(\langle\mathbf{B}_l,\mathbf{X}_i\rangle) = \mathbb{E}_{\underline{\mathbf{X}}^c}(\langle\mathbf{B}_{l'},\mathbf{X}_i\rangle) = 0$. We are left with:

$$\begin{aligned}
&\mathbb{E}_{\underline{\mathbf{X}}^c} D_{KL}(f_l(\mathbf{y}|\underline{\mathbf{X}}^c)||f_{l'}(\mathbf{y}|\underline{\mathbf{X}}^c) \\
&= \sum_{i\in[n]} \mathbb{E}_{\mathbf{X}}\left[\sigma_l \cdot \langle\mathbf{X}_i,\mathbf{B}_l\rangle - \sigma_l\langle\mathbf{X}_i,\mathbf{B}_{l'}\rangle\right] \\
&= \sum_{i\in[n]} \mathbb{E}_{\mathbf{X}}\left[\sigma_l \cdot \langle\mathbf{X}_i,\mathbf{B}_l - \mathbf{B}_{l'}\rangle\right] \\
&= \sum_{i\in[n]} \mathbb{E}_{\mathbf{X}}\left[\frac{\langle\mathbf{X}_i,\mathbf{B}_l - \mathbf{B}_{l'}\rangle}{1+\exp^{-\langle\mathbf{X}_i,\mathbf{B}_l\rangle}}\right] \\
&\leq n\mathbb{E}_{\mathbf{X}}\left[|\langle\text{vec}(\mathbf{X}_i),\text{vec}(\mathbf{B}_l - \mathbf{B}_{l'})\rangle|\right]. \tag{60}
\end{aligned}$$

483

Define the random variable $\widetilde{X} \triangleq \langle \mathbf{X}_i, \mathbf{B}_l - \mathbf{B}_{l'} \rangle$. $\widetilde{X}$ is a normally distributed random variable with mean $\mu_{\{\widetilde{X}\}} = 0$ and variance $\sigma_{\widetilde{X}}^2 = \sigma^2 \text{vec}(\mathbf{B}_l - \mathbf{B}_{l'})^T \mathbf{I}_{m_1 m_2} \text{vec}(\mathbf{B}_l - \mathbf{B}_{l'})$. Now, define the half-Normal random variable $\bar{X} \triangleq |\widetilde{X}|$. $\bar{X}$ is a half-Normal distribution with mean:

$$
\mathbb{E}_{\bar{X}}[\bar{X}] = \sigma \sqrt{\text{vec}(\mathbf{B}_l - \mathbf{B}_{l'})^T \text{vec}(\mathbf{B}_l - \mathbf{B}_{l'})} \sqrt{\frac{2}{\pi}}
$$

$$
= \sigma \|\mathbf{B}_l - \mathbf{B}_{l'}\|_F \sqrt{\frac{2}{\pi}}. \tag{61}
$$

Plugging in (61) and (60) into (58) gives us,

$$
\mathbb{I}(\mathbf{y}; l | \underline{\mathbf{X}}^c) \leq n\sigma \|\mathbf{B}_l - \mathbf{B}_{l'}\|_F \sqrt{\frac{2}{\pi}} \overset{(f)}{\leq} n\sigma \sqrt{\frac{4r}{r-1}} \sqrt{\frac{2}{\pi}} \varepsilon,
$$

where (f) follows from (57). ∎

REFERENCES

[1] R. E. Wright, "Logistic regression." in *Reading and Understanding Multivariate Statistics*. American Psychological Association, 1995, pp. 217–244.

[2] J. P. Dumas, M. A. Lodhi, B. A. Taki, W. U. Bajwa, and M. C. Pierce, "Computational endoscopy—a framework for improving spatial resolution in fiber bundle imaging," *Optics Letters*, vol. 44, no. 16, pp. 3968–3971, 2019.

[3] M. A. Lodhi and W. U. Bajwa, "Learning product graphs underlying smooth graph signals," *arXiv preprint arXiv:2002.11277*, 2020.

[4] H. Hung and C.-C. Wang, "Matrix variate logistic regression model with application to EEG data," *Biostatistics*, vol. 14, no. 1, pp. 189–202, 2013.

[5] J. Zhang and J. Jiang, "Rank-optimized logistic matrix regression toward improved matrix data classification," *Neural Computation*, vol. 30, no. 2, pp. 505–525, 2018.

[6] L. P. Barnes and A. Ozgur, "Minimax bounds for distributed logistic regression," *arXiv preprint arXiv:1910.01625*, 2019.

[7] A. R. Zhang, Y. Luo, G. Raskutti, and M. Yuan, "ISLET: Fast and optimal low-rank tensor regression via importance sketching," *SIAM J. on Mathematics of Data Science*, vol. 2, no. 2, pp. 444–479, 2020.

[8] G. Raskutti, M. Yuan, and H. Chen, "Convex regularization for high-dimensional multiresponse tensor regression," *The Annals of Statstics*, vol. 47, no. 3, pp. 1554–1584, 2019.

[9] D. J. Foster, S. Kale, H. Luo, M. Mohri, and K. Sridharan, "Logistic regression: The importance of being improper," in *Proc. Conf. On Learning Theory (PMLR)*, 2018, pp. 167–208.

[10] F. Abramovich and V. Grinshtein, "High-dimensional classification by sparse logistic regression," *IEEE Trans. on Information Theory*, vol. 65, no. 5, pp. 3068–3079, 2018.

[11] J. V. Shi, Y. Xu, and R. G. Baraniuk, "Sparse bilinear logistic regression," *arXiv preprint arXiv:1404.4104*, 2014.

[12] B. An and B. Zhang, "Logistic regression with image covariates via the combination of ℓ_1 and Sobolev regularizations," *PLOS One*, vol. 15, no. 6, p. e0234975, 2020.

[13] Q. Berthet and N. Baldin, "Statistical and computational rates in graph logistic regression," in *Proc. Int. Conf. on Artificial Intelligence and Statistics (PMLR)*, 2020, pp. 2719–2730.

[14] A. Jung, Y. C. Eldar, and N. Görtz, "On the minimax risk of dictionary learning," *IEEE Trans. on Inf. Theory*, vol. 62, no. 3, pp. 1501–1515, 2016.

[15] Z. Shakeri, W. U. Bajwa, and A. D. Sarwate, "Minimax lower bounds for kronecker-structured dictionary learning," in *Proc. 2016 IEEE Int. Symp. on Inf. Theory (ISIT)*. IEEE, 2016, pp. 1148–1152.

[16] B. Yu, "Assouad, Fano, and Le Cam," in *Festschrift for Lucien Le Cam*. Springer, 1997, pp. 423–435.

[17] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Review*, vol. 51, no. 3, pp. 455–500, 2009.

[18] S. Negahban and M. J. Wainwright, "Restricted strong convexity and weighted matrix completion: Optimal bounds with noise," *The J. of Machine Learning Res. (JMLR)*, vol. 13, pp. 1665–1697, 2012.

[19] R. Z. Khas'minskii, "A lower bound on the risks of non-parametric estimates of densities in the uniform metric," *Theory of Probability & Its Applications*, vol. 23, no. 4, pp. 794–798, 1979.

[20] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, 2012.

[21] D. P. Dubhashi and A. Panconesi, *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge University Press, 2009.

[22] Z. Shakeri, W. U. Bajwa, and A. D. Sarwate, "Minimax lower bounds on dictionary learning for tensor data," *IEEE Trans. on Inf. Theory*, vol. 64, no. 4, pp. 2706–2726, 2018.

[23] M. J. Wainwright, "Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting," *IEEE Trans. on Inf. Theory*, vol. 55, no. 12, pp. 5728–5741, 2009.