Minimax risk of learning low-rank multinomial distributions

Arvind Rathnashyam
RPI Math, rathna@rpi.edu

Mohammad Mohammadi Amiri
RPI CS, mamiri@rpi.edu

Alex Gittens
RPI CS, gittea@rpi.edu

We derive bounds on the minimax risk of estimating discrete multinomial models in the total variation distance from n i.i.d one-hot samples when the underlying probability matrix $X \in \mathbb{R}^{k\times d}$ arises from a low-rank latent structure. Specifically, we consider $X(H) = \exp(H)/\sum_{i,j}\exp(H_{i,j})$ where $\text{rank}(H) \ll \min\{k,d\}$. We use Fano's inequality to prove that any estimator incurs $\Omega(\sqrt{r(d+k)/n})$ error. To establish achievability, we analyze a nuclear norm penalized maximum log-likelihood estimator to prove a near-optimal error of $O(\sqrt{r(d+k)\log(d+k)/n})$ over the problem space. This is an improvement over the standard $O(\sqrt{kd/n})$ bound obtained by using the empirical estimator.

1 Introduction

Real world data often admits a low-rank structure, particularly in applications involving co-occurrence or contingency tables. In large language models, the word context probabilities often admits a low-rank representation. The probability of observing (i, j) for $i \le k$ and $j \le d$ is given by

$$X_{i,j} = \frac{1}{Z} \exp\left((UV^\top)_{i,j}\right), \quad Z = \sum_{i=1}^{k} \sum_{j=1}^{d} \exp((UV^\top)_{i,j}),$$

where $U \in \mathbb{R}^{k \times r}$, $V \in \mathbb{R}^{d \times r}$ where $r \ll \min\{k, d\}$. In the unstructured case when there is not a low-rank assumption on X, the error scales like $\sqrt{kd/n}$ where n is the number of i.i.d samples from X. In this work, we precisely quantify the improvements of error rate when the underlying matrix has low-rank structure. We also develop algorithms that are able to take advantage of this extra structure and obtain near-optimal error.

Contributions.

1. We provide a information-theoretic lower bound on the worst case error for any estimator $\hat{X}$. We show that any estimator incurs at least $\Omega(\sqrt{r(d+k)/n})$ error in the total variation distance by constructing a packing set and using Fano's inequality.
2. We establish near-achievability by constructing a nuclear norm penalized maximum likelihood estimator that achieves error at most $O(\sqrt{r(d+k)\log(d+k)/n})$ worst case error.

Implications. The main implication of our work is that any algorithm requires $\Omega(r(d+k)/\varepsilon^2)$ samples to obtain a $O(\varepsilon)$ estimation error with high probability. Furthermore we show the existence of an algorithm that requires $O(r(d+k)\log(d+k)/\varepsilon^2)$ samples to obtain a $O(\varepsilon)$ estimation error. We also make similar conclusions for learning multinomial distributions that can be parameterized by low-rank matrices.

1.1 Applications

Our results apply to problems with observing samples that are two-dimensional.

- **Word embeddings and language modeling.** An application of our work is the learning of token co-occurence models, similar to word2vec [GL14, Mik13]. In this setting, the probability of observing a token (i, j) is proportional to the softmax output of low-rank matrix $UV^\top$. Our work directly shows the number of samples needed to estimate the distribution with high probability.
- **Contingency tables.** In statistics, contingency tables model the frequencies for two categorical variables represented in a matrix [Eve92]. When the underlying distribution is low-rank, we show that we require less samples to obtain a high accuracy estimation for the distribution.
- **Recommendation systems.** User-item interactions are often modeled by a low-rank probability matrix [Hof04]. Observing a user-item interaction is like sampling (i, j) from a multinomial distribution. The underlying probability matrix modeling the interaction probabilities is a low-rank multinomial estimation problem.

1.2 Roadmap

In Section 3, we discuss the model classes and observation model we study in detail. We study the minimax risk for learning multinomial distributions and low-rank multinomial distributions (Appendix A) in the total variation distance (cf. Equation (1)). In Proposition 14, we show the sub-optimality of Le Cam's method, only being able to obtain a lower bound of $\mathfrak{R}^* = \Omega(\sqrt{1/n})$ for learning multinomial distributions. We then obtain the optimal lower bound with Fano's method in Theorem 1 of $\mathfrak{R}^* \ge \Omega(\sqrt{kd/n})$. In Theorem 17, we obtain an upper bound on the minimax risk in the total variational distance for multinomial distributions,

we use the empirical estimator to obtain $\mathfrak{R}^* \leq O(\sqrt{kd/n})$. In Theorem 18, we study the lower bound on the minimax risk of learning low-rank multinomial distributions in the total variational distance, with Fano's inequality we obtain $\mathfrak{R}^* = \Omega(\sqrt{r(d+k-r)/n})$. In Theorem 22, we derive an upper bound on the minimax risk of learning low-rank word2vec models in the total variational distance using a nuclear norm penalized maximum likelihood estimator and obtain $\mathfrak{R}^* = O(\sqrt{r(d+k)\log(d+k)/n})$.

2 Background and related works

In this section, we review the essential concepts from probability theory, linear algebra, and learning theory for our analysis. We then discuss the place of our contributions among the related work in the literature.

2.1 Probability theory preliminary

We will first define metrics for measuring distance between discrete probability distributions. Suppose Ω is a finite event space. The Kullback-Leibler (KL) divergence [KL51] for a discrete distribution is defined as

$$D_{\mathrm{KL}}(P \parallel Q) \coloneqq \sum\nolimits_{x\in\Omega} P(x) \log\left(\frac{P(x)}{Q(x)}\right).$$

For discrete distributions, the total variation distance is

$$\|P - Q\|_{\mathrm{TV}} \coloneqq \frac{1}{2}\sum\nolimits_{x\in\Omega} |P(x) - Q(x)|. \tag{1}$$

The KL divergence and total variation distance are connected by Pinsker's inequality [CK11],

$$\mathrm{TV}(P, Q) \leq \sqrt{2D_{\mathrm{KL}}(P \parallel Q)}, \tag{2}$$

which implies that if two distributions are close in the KL divergence, then they are close in the total variation distance. With Pinsker's inequality, we are able to reduce total variation upper bounds to an upper bound on the KL divergence. In this work we focus on learning multinomial distributions. Given n i.i.d. samples from a multinomial distribution parameterized by a vector $\mathbf{x} \in \Delta^{k-1}$, the probability of obtaining a count vector $\mathbf{y} \in \mathbb{N}^{d-1}$ is

$$\Pr_{\mathbf{y}\sim\mathbf{x}}(\mathbf{y}) = \frac{n!}{y_1! \cdots y_k!} x_1^{y_1} \cdots x_k^{y_k}. \tag{3}$$

2.2 Linear algebra preliminary

For a matrix $A \in \mathbb{R}^{n\times m}$, the Schatten-$s$ norm [Bha13] for $s \geq 1$ is defined as

$$\|A\|_{\sigma,s} = \left(\sum\nolimits_k \sigma_k^s(A)\right)^{1/s},$$

where $\sigma_k(A)$ are the singular values of A in descending order. The spectral norm is given as $\|A\|_{\sigma,\infty}$ and the nuclear norm is given as $\|A\|_{\sigma,1}$. The nuclear norm is a convex envelope for the rank of a matrix on the spectral norm ball [Faz02]. It is therefore a popular choice as a regularization method for problems with low-rank constraints or assumptions (see e.g. [RFP10, KLT11, DPVDBW14]). For the set of matrices in $\mathbb{R}^{k\times d}$, we define the scalar dot product $\langle A, B\rangle = \mathrm{tr}(A^\top B)$. The singular value decomposition for $A \in \mathbb{R}^{m\times n}$ is given as $U\Sigma V^\top$ where $U \in \mathbb{R}^{m\times \mathrm{rank}(A)}$, $V \in \mathbb{R}^{n\times \mathrm{rank}(A)}$ are tall, semi-orthogonal matrices and $\Sigma \in \mathbb{R}^{\mathrm{rank}(A)\times \mathrm{rank}(A)}$ is a PSD diagonal matrix. The projection of a matrix A is defined as $\mathcal{P}_A \coloneqq UU^\top$ where $A = U\Sigma V^\top$ is the singular value decomposition. For a vector $\mathbf{x}$, we use $\|\mathbf{x}\|_0$ to denote the number of nonzero entries, and the Hamming distance to measure coordinate-wise mismatches.

2.3 Learning theory preliminary

The minimax risk is defined as the minimal worst-case error achievable by any estimator. For the total variation distance, it is defined as

$$\mathfrak{R}^* \coloneqq \min_{\hat{X}} \max_{X \in \Theta} \|\hat{X} - X\|_{\mathrm{TV}}, \tag{4}$$

where Θ is the parameter space A lower bound on the minimax risk of $R^* \geq \Omega(a)$ implies that any algorithm obtains at least $\Omega(a)$ risk for some problem instance. An upper bound on the minimax risk of $R^* \leq O(a)$ implies the existence of an algorithm obtaining at most $O(a)$ risk on any instance of the problem. Minimax lower bounds are often proved with reductions to hypothesis testing problems. Le Cam's two-point method [LC12] reduces the problem to a binary hypothesis testing problem between two chosen elements in Θ and Fano's inequality [Cov99] reduces the problem to a multi-way hypothesis testing problem between elements in a packing set of Θ.

2.4 Related works

For multinomial distributions, the minimax risk in the total variational distance without any structural constraints has been characterized as $\Theta(\sqrt{d/n})$ in [HJW15]. In particular, Han et al. [HJW15] show that the empirical estimator obtains the minimax optimal risk. In our work, we also show that for multinomials that can be parameterized by a vector that the minimax error is $\sqrt{d/n}$ and is achievable by the empirical estimator. However, we show that the empirical estimator does not achieve the minimax optimal rate when the multinomial can be parameterized by a low-rank matrix. Another similar problem in the literature is to the minimax risk of the learning the total variation distance, $\|P-Q\|_1$, where Q is known and P is unknown, and one obtains n i.i.d. samples from P. Jiao et al. [JHW18] show that the empirical estimator obtains an error rate of $\sqrt{d/n}$ in the high data regime.
The problem of estimating structured probability distributions from observations has been studied in the literature in the context of low-rank matrix completion [CP09, KMO09]. The observation models typically have additive Gaussian noise or a Bernoulli sampling model. We consider the multinomial sampling model. Davenport et al. [DPVDBW14] consider a nuclear norm constrained minimization algorithm for low-rank matrix completion. Lafond et al. [LKMS14] study a similar nuclear norm penalized algorithm to improve upon their results. We take inspiration from these algorithms and develop a nuclear norm penalized maximum likelihood algorithm to establish near achievability of our minimax lower bounds.

3 Model classes and observation model

In this section, we first discuss the model classes we will be studying. We then establish the minimax rates for learning low-rank multinomial distributions under the total variational distance. Our strategy is to first derive information-theoretic lower bounds with Fano's method by constructing a packing set of multinomial distributions. We then show near-optimal achievability using a nuclear-norm penalized maximum log likelihood estimator.

3.1 Observation model

We observe n i.i.d. one-hot samples $Y_1, \ldots, Y_n \in \{0,1\}^{k \times d}$ such that $Y_i = E_{u,v}$ where $(E_{u,v})_{a,b} = \mathbf{1}\{u = a, v = b\}$. Each sample is distributed as $\mathbf{Pr}\{Y_i = E_{u,v}\} = X_{u,v}$ for an $X \in \Delta^{kd-1}$, where Δ^{kd-1} represents the probability simplex over $\mathbb{R}^{k \times d}$. The empirical sampling matrix is defined as $Y \coloneqq \frac{1}{n}\sum_i Y_i$, and is an unbiased estimator of X.

3.2 Model classes and main results

We now introduce the three hypothesis results that have different levels of structural complexity. Each hypothesis class is motivated by practical examples.

(M1) Unstructured multinomial. The parameter set $\Theta_1 = \{X \in \mathbb{R}^{k\times d} : X \in \Delta^{kd-1}\}$, represents the most general set of probability matrices without any structural constraints.

Theorem 1. *Let $\Theta = \{X \in \mathbb{R}^{k\times d} : X \in \Delta^{kd-1}\}$. There exists an absolute constants $c, C > 0$ such that*

$$c\sqrt{kd/n} \le \min_{\hat{X}} \max_{X\in\Theta} \|\hat{X} - X\|_{\mathrm{TV}} \le C\sqrt{kd/n},$$

with probability at least $1/2$.

Our results in Theorem 1 agree with the findings in Han et al. [HJW15]. Our proof of the upper bound is a simple mean deviation argument with Chebyshev's inequality when we use the empirical estimator. For our proof of the lower bound, we use a standard Fano style argument by constructing a packing set of Θ. Interestingly, Le Cam's method does not obtain the optimal lower bound.

Proposition 2. *Let $\Theta = \{X \in \mathbb{R}^{k\times d} : \sum_{i,j} X_{i,j} = 1, X_{i,j} \ge 0\}$. Then the minimax lower bound obtained by Le Cam's method is $\Omega(\sqrt{1/n})$.*

We defer the technical details of Proposition 14 to the appendix.

(M2) Low-rank probability matrix. The hypothesis $\Theta_2 = \{X \in \mathbb{R}^{k\times d} : X \in \Delta^{kd-1}, \mathrm{rank}(X) \le r\}$, where $r \ll \min\{k, d\}$ incorporates our first layer of structural constraints by requiring the probability matrices to have low-rank. We can note that $\Theta_2 \subset \Theta_1$ which gives us reason to hypothesize that the minimax error over Θ_2 should be less than the minimax error of Θ_1, considering the effective dimensionality of the underlying distribution is reduced to $O(r(k+d))$. To ensure stable estimation, we require a regularity condition:

Assumption 3. *There exists a constant $\nu \ge 1$ such that*

$$\max\left\{\max_{1\le i\le k}\sum\nolimits_j X_{i,j}, \max_{1\le j\le d}\sum\nolimits_i X_{i,j}\right\} \le \nu/\min(k,d).$$

Assumption 3 requires no row or column to be sampled with too high probability. These assumptions are standard in the low-rank matrix completion problem [FSSS11, LKMS14, Klo14].

Theorem 4. *Let $\Theta = \{X \in \mathbb{R}^{k\times d} : \sum_{i,j} X_{i,j} = 1, \mathrm{rank}(X) \le r\}$. There exists an absolute constant $c > 0$ such that*

$$\min_{\hat{X}} \max_{X\in\Theta} \|\hat{X} - X\|_{\mathrm{TV}} \ge c\sqrt{r(k+d)/n},$$

with probability at least $1/2$.

Theorem 5. *Let $\Theta = \{X \in \mathbb{R}^{k\times d} : \sum_{i,j} X_{i,j} = 1, \mathrm{rank}(X) \le r\}$ and suppose Θ satisfies Assumption 3. There exists an absolute constant $C > 0$ such that*

$$\min_{\hat{X}} \max_{X\in\Theta} \|\hat{X} - X\|_{\mathrm{TV}} \le C\sqrt{r(k+d)\log(k+d)/n},$$

with probability at least $1/2$.

Our proof of Theorem 5 is by construction. The nuclear-norm penalized maximum likelihood estimator referenced in Section 4.2 achieves the lower bound up to a logarithmic factor.

(M3) Low-rank underlying softmax. This model class directly models the parameterization used in popular word embedding algorithms such as word2vec, where the conditional probabilities are proportional to the softmax of the low-dimensional vector representations:

$$\mathbf{Pr}_\theta(c \mid w) \;=\; \frac{\exp\left(u_w^\top v_c\right)}{\sum_{c'=1}^{d}\exp\left(u_w^\top v_{c'}\right)},$$

where u_w, v_c represent the vector embeddings. We define Θ_3 as the matrix softmax of a low-rank matrix.

Theorem 6. *Let* $\Theta = \{X \in \mathbb{R}^{k\times d} : \sum_{i,j} X_{i,j} = 0, \mathrm{rank}(X) \le r\}$. *There exists an absolute constant* $c > 0$ *such that*

$$\min_{\hat{X}} \max_{X\in\Theta} \|f(\hat{X}) - f(X)\|_{\mathrm{TV}} \ge c\sqrt{r(k+d)/n},$$

with probability at least $1/2$.

Assumption 7. *There exists a* $\mu > 0$ *such that* $f(X)_{i,j} \ge \mu/kd$ *for all* $i \le k$ *and* $j \le d$.

Theorem 8. *Let* $\Theta = \{X \in \mathbb{R}^{k\times d} : \sum_{i,j} X_{i,j}\mathrm{rank}(X) \le r\}$. *There exists an absolute constant* $C > 0$ *such that*

$$\min_{\hat{X}} \max_{X\in\Theta} \|f(\hat{X}) - f(X)\|_{\mathrm{TV}} \le C\sqrt{r(k+d)\log(k+d)/n},$$

with probability at least $1/2$.

Despite the nonlinearity of the softmax function, the statistical complexity of the the problem remains $O(\sqrt{r(k+d)})$, similar to the low-rank matrix case in Theorem 4.

4 Proofs of main results

In this section we will go through the proofs of our main results. We defer the technical details to the appendix and use lemmas to give a higher-level overview of the proof.

4.1 Lower bound overview

Our goal is to establish information-theoretic lower bounds on the minimax risk. We achieve this by first reducing the the estimation problem to a multi-hypothesis testing problem [Tsy08, SC21]. The main idea of our approach is to construct a finite packing set of the parameter space Θ that are well separated in the total variation distance. In particular we form the set $\{X^{(1)}, \dots, X^{(M)}\} \subset \Theta$ such that for distinct $u, v \in \{1, \dots, M\}$,

$$\|X^{(u)} - X^{(v)}\|_{\mathrm{TV}} \ge \varepsilon.$$

Following the results of Scarlett & Cevher [SC21, Corollary 1], the construction of the packing set implies

$$\mathbf{Pr}\left(\min_{\hat{X}} \max_{X\in\Theta} \|\hat{X} - X\|_{\mathrm{TV}} \ge \varepsilon/2\right) \ge \min_{\tilde{X}} \mathbf{Pr}(\tilde{X} \ne B), \tag{5}$$

where $B \in \mathbb{R}^{k\times d}$ is uniformly distributed over the packing set, $\{X^{(1)}, \dots, X^{(M)}\}$. To lower bound the probability of a mismatch, we apply Fano's inequality [Cov99]:

$$\mathbf{Pr}(\widetilde{X} \ne B) \ge 1 - \frac{\mathcal{I}(B;Y) + \log 2}{\log M(\varepsilon, \Theta)}. \tag{6}$$

Thus the proof strategy reduces to constructing a packing set such that the following three conditions hold:

1. Well separated. $\|X^{(u)} - X^{(v)}\| \ge \Omega(\varepsilon)$ for all distinct u, v.
2. Small mutual information. The mutual information between the true parameter, B, and the observations Y, is sufficiently small, i.e. $\mathcal{I}(B;Y) = O(n\varepsilon^2)$.
3. Large packing set. We require the packing set to be exponentially large, i.e $\log M(\varepsilon, \Theta) = \Omega(r(k+d))$ in the structured case.

Then applying Equation (5) and Equation (6) yields the desired minimax lower bound.

4.2 Upper bound overview

To establish matching minimax upper bounds, we construct algorithms and control the worst-case error over Θ. From the definition of the minimax risk (see Equation (4)), we note that

$$\min_{\hat{X}} \max_{X \in \Theta} \mathbf{E} \, \|X - \hat{X}\|_{\mathrm{TV}} \leq \max_{X \in \Theta} \mathbf{E} \, \|X - \hat{X}\|_{\mathrm{TV}}.$$

We consider two separate algorithms:

- Unstructured case. When the underlying matrix has no low-rank structure (even in a latent space), we prove that the empirical estimator achieves the minimax optimal rate.
- Structured case. When the underlying matrix has a low-rank structure (or low-rank in a latent space), we propose a nuclear norm penalized maximum log likelihood estimator inspired by [LKMS14].

4.3 Low rank underlying softmax distribution

In this section we give a high level overview of our minimax bounds for learning low-rank underlying softmax distributions. We defer technical details to the appendix and use lemmas we prove in the appendix to give the sketch.

Proof of Theorem 6. Packing set construction. Let $m := (r-1)(k+d+(r-1/2)/2$ represent the number of active entries in X that lie in the top $(r-1)/2$ rows or left $(r-1)/2$ rows. We select a binary code $\mathcal{V} \subset \{0,1\}^m$ with minimum Hamming distance $\Omega(m)$. We can show that $\mathcal{V}$ exists such that $M \geq 2^{\Omega(m)}$ using the Gilbert-Varshamov (GV) bound. For each codeword $u \in \mathcal{V}$, we define

$$X_{i,j}^{(u)} = \begin{cases} (2c_{\pi(i,j)} - 1)\sqrt{m/n}, & i \leq (r-1)/2 \text{ or } j \leq (r-1)/2 \\ -\xi, & i > (r-1)/2 \text{ and } j > (r-1)/2, \end{cases} \tag{7}$$

where $\pi(i,j)$ is a mapping function from the matrix position to the element in the code. Using a block partion of $X^{(u)}$, we can observe that $\mathrm{rank}(X^{(u)}) \leq r$ for all $u \leq M$.

Lemma 9. *Consider the packing set of Θ with size M formed in Equation (7). Then for any $u, v \leq M$ s.t. $u \neq v$, $\|X^{(u)} - X^{(v)}\|_{\mathrm{TV}} = \Omega(\sqrt{m/n})$.*

Lemma 9 can be observed from noting that the Hamming distance between any $X^{(u)}$ and $X^{(v)}$ in the packing set is $\Omega(m)$, and the element-wise difference in the total variation distance is at least $\Omega(\sqrt{m/n})$.

Lemma 10. *Consider the packing set of Θ with size M formed in Equation (7). Then for any $u, v \leq M$ s.t. $u \neq v$, the KL divergence between $X^{(u)}$ and $X^{(v)}$ satisfies $D_{\mathrm{KL}}(f(X^{(u)}) \parallel f(X^{(v)})) = O(m/n)$.*

Combining Lemmas 9 and 10 with Pinsker's inequality (see Equation (2)) reveals that our packing set construction achieves the information-theoretic optimal separation. The separation in the packing set matches the upper bound implied by Pinsker's inequality. We can bound the mutual information by $I(B;Y) \leq \max_{u,v} D_{\mathrm{KL}}(X^{(u)} \parallel X^{(v)})$. Then from Lemma 10, we find that $I(B;Y) = O(m)$. We can apply Fano's inequality to the M-ary hypothesis testing problem,

$$\min_{\hat{X}} \max_{X \in \Theta} \|\hat{X} - f(X)\|_{\mathrm{TV}} \geq 1 - \frac{Cm + \log 2}{\log M}.$$

Then noting that $\log M = \Omega(m)$ we find the probability that $\|\hat{X} - f(X)\|_{\mathrm{TV}} = \Omega(\sqrt{m/n})$ with probability at least $1/2$. ■

Proof of Theorem 8. Consider the set of matrices, $G = \{M \in \mathbb{R}^{k \times d} : \max_{i,j} |M_{i,j}| \leq C, \sum_{i,j} M_{i,j} = 0\}$, where C is a small constant that ensures that Assumption 7 is satisfied. We define the nuclear norm penalized maximum log likelihood estimator as

$$\hat{X} = \arg\min_{M \in G} \{-\Phi_Y(f(M)) + \lambda \|M\|_{\sigma,1}\},$$

where the loss is defined as $\Phi_Y(M) \coloneqq \sum_{i,j} Y_{i,j} \log(f(M))_{i,j}$. The nuclear norm penalization is a popular choice in the literature as the nuclear norm is a convex envelope of the rank. Our motivation is from the nuclear norm constrained algorithm presented by Davenport et al. [DPVDBW14], and the nuclear norm constrained algorithm presented by Lafond et al. [LKMS14]. We first consider the choice for regularization parameter λ.

Lemma 11. *For X to be in the $\arg\min_{M \in G}\{-\Phi_Y(f(M)) + \lambda\|M\|_{\sigma,1}\}$, then the regularization parameter must satisfy $\lambda = \Omega(\sqrt{\log(k+d)/(n\min(k,d))})$.*

The proof of Lemma 11 revolves around the KKT conditions of the optimization problem and bounding $\|Y - f(X)\|_{\sigma,\infty}$ with matrix Bernstein. From the minimization guarantee, for any $M \in G$,

$$-\Phi_Y(f(\hat{X})) + \lambda\|\hat{X}\|_{\sigma,1} \leq -\Phi_Y(f(M)) + \lambda\|M\|_{\sigma,1}. \tag{8}$$

Lemma 12. *For any two matrices, A, B,*

$$\|A\|_{\sigma,1} - \|B\|_{\sigma,1} \leq \sqrt{\operatorname{rank}(A)}\|A - B\|_{\sigma,1}.$$

We defer the proof of Lemma 12 to Appendix B. We can then rearrange Equation (8), and obtain

$$\Phi_Y(f(M)) - \Phi_Y(f(\hat{X})) \leq \lambda\left(\|M\|_{\sigma,1} - \|\hat{X}\|_{\sigma,1}\right) \leq \lambda\sqrt{\operatorname{rank}(M)}\|M - \hat{X}\|_{\sigma,1},$$

where we use Lemma 12 in the second inequality.

Lemma 13. *Suppose f is the softmax function applied over the whole matrix. There exists a constant, C dependent on the set G, such that $\|A - B\|_{\sigma,1} \leq C\sqrt{kd}\|f(A) - f(B)\|_{\sigma,1}$.*

Using the guarantee in Lemma 13,

$$\lambda\sqrt{\operatorname{rank}(M)}\|M - \hat{X}\|_{\sigma,1} \leq \lambda\sqrt{kd\operatorname{rank}(M)}\|f(M) - f(\hat{X})\|_{\mathrm{TV}} \leq \lambda\sqrt{kd\operatorname{rank}(M)D_{\mathrm{KL}}(f(M) \parallel f(\hat{X}))}.$$

In the above, in the second inequality we used Pinsker's inequality (see Equation (2)). From our formulation of the loss function, we find that for $M \in G$, $D_{\mathrm{KL}}(f(M) \parallel f(\hat{X})) = \mathbf{E}[\Phi(f(M)) - \Phi(f(\hat{X}))]$. After applying Markov's inequality, we find that with high probability

$$D_{\mathrm{KL}}(f(M) \parallel f(\hat{X})) \leq \lambda\sqrt{kd\operatorname{rank}(M)}\|f(M) - f(\hat{X})\|_{\mathrm{TV}} \leq \lambda\sqrt{kd\operatorname{rank}(M)D_{\mathrm{KL}}(f(M) \parallel f(\hat{X}))}.$$

Then, we can apply Pinsker's inequality once more to lower bound the LHS after rearranging,

$$\|f(M) - f(\hat{X})\|_{\mathrm{TV}} \leq \lambda\sqrt{kd\operatorname{rank}(M)},$$

then from our choice of λ in Lemma 11, our proof is complete. ■

5 Discussion

In this paper, we have derived minimax optimal rates for learning structured multinomial distributions in the total variation distance. Our results reveal a decrease in the minimax error from $\Omega(\sqrt{kd/n})$ in the structured case to $\Omega(\sqrt{r(d+k)/n})$. We develop a nuclear-norm penalized maximum log likelihood estimator to show that the lower bound is achievable within logarithmic factors. We also find that the advantages of low-rank structure our preserved under the softmax function, providing some insight into the theoretical basis of modern representation learning.

Future work can be to to develop matching algorithms for the lower bound and remove the logarithmic factor that exists for the current algorithms we have developed in this paper. Another interesting idea is to incorporate the idea of sequential dependence, where the probability of observing sample x_t is dependent on samples $x_{<t}$. Studying the minimax optimal rates and developing matching algorithms would be a promising direction for future research.

References

[Bha13] Rajendra Bhatia. *Matrix analysis*, volume 169. Springer Science & Business Media, 2013.

[BSS06] Mokhtar S Bazaraa, Hanif D Sherali, and Chitharanjan M Shetty. *Nonlinear programming: theory and algorithms*. John wiley & sons, 2006.

[CK11] Imre Csiszár and János Körner. *Information theory: coding theorems for discrete memoryless systems*. Cambridge University Press, 2011.

[Cov99] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[CP09] Emmanuel J Candes and Yaniv Plan. Accurate low-rank matrix recovery from a small number of linear measurements. In *2009 47th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1223–1230. IEEE, 2009.

[DPVDBW14] Mark A Davenport, Yaniv Plan, Ewout Van Den Berg, and Mary Wootters. 1-bit matrix completion. *Information and Inference: A Journal of the IMA*, 3(3):189–223, 2014.

[Eve92] Brian S Everitt. *The analysis of contingency tables*. CRC Press, 1992.

[Faz02] Maryam Fazel. *Matrix rank minimization with applications*. PhD thesis, PhD thesis, Stanford University, 2002.

[Fel91] William Feller. *An introduction to probability theory and its applications, Volume 2*, volume 2. John Wiley & Sons, 1991.

[Fir60] William J Firey. Remainder formulae in taylor's theorem. *The American Mathematical Monthly*, 67(9):903–905, 1960.

[FSSS11] Rina Foygel, Ohad Shamir, Nati Srebro, and Russ R Salakhutdinov. Learning with the weighted trace-norm under arbitrary sampling distributions. *Advances in neural information processing systems*, 24, 2011.

[Gil52] Edgar N Gilbert. A comparison of signalling alphabets. *The Bell system technical journal*, 31(3):504–522, 1952.

[GL14] Yoav Goldberg and Omer Levy. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*, 2014.

[HJ94] Roger A Horn and Charles R Johnson. *Topics in matrix analysis*. Cambridge university press, 1994.

[HJW15] Yanjun Han, Jiantao Jiao, and Tsachy Weissman. Minimax estimation of discrete distributions. In *2015 IEEE International Symposium on Information Theory (ISIT)*, pages 2291–2295. IEEE, 2015.

[Hof04] Thomas Hofmann. Latent semantic models for collaborative filtering. *ACM Transactions on Information Systems (TOIS)*, 22(1):89–115, 2004.

[HTW15] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. Statistical learning with sparsity. *Monographs on statistics and applied probability*, 143(143):8, 2015.

[JHW18] Jiantao Jiao, Yanjun Han, and Tsachy Weissman. Minimax estimation of the $l\_\{1\}$ distance. *IEEE Transactions on Information Theory*, 64(10):6672–6706, 2018.

[Kar39] William Karush. Minima of functions of several variables with inequalities as side constraints. *M. Sc. Dissertation. Dept. of Mathematics, Univ. of Chicago*, 1939.

[KL51] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[Klo14] Olga Klopp. Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, 20(1), February 2014.

[KLT11] Vladimir Koltchinskii, Karim Lounici, and Alexandre B Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. 2011.

[KMO09] Raghunandan Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from noisy entries. *Advances in neural information processing systems*, 22, 2009.

[KT13] Harold W Kuhn and Albert W Tucker. Nonlinear programming. In *Traces and emergence of nonlinear programming*, pages 247–258. Springer, 2013.

[LC12] Lucien Le Cam. *Asymptotic methods in statistical decision theory*. Springer Science & Business Media, 2012.

[LKMS14] Jean Lafond, Olga Klopp, Eric Moulines, and Joseph Salmon. Probabilistic low-rank matrix completion on finite alphabets. *Advances in Neural Information Processing Systems*, 27, 2014.

[M+89] Colin McDiarmid et al. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

[Mik13] Tomas Mikolov. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 3781, 2013.

[RFP10] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[SC21] Jonathan Scarlett and Volkan Cevher. An introductory guide to fano's inequality with applications in statistical estimation. *Information-Theoretic Methods in Data Science*, pages 487–528, 2021.

[Tro12] Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12:389–434, 2012.

[Tsy08] Alexandre B Tsybakov. Nonparametric estimators. In *Introduction to Nonparametric Estimation*, pages 1–76. Springer, 2008.

[Var57] Rom Rubenovich Varshamov. Estimate of the number of signals in error correcting codes. *Docklady Akad. Nauk, SSSR*, 117:739–741, 1957.

[YB99] Yuhong Yang and Andrew Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, pages 1564–1599, 1999.

[Yu97] Bin Yu. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam: research papers in probability and statistics*, pages 423–435. Springer, 1997.

A Proofs

In this section, we provide the proofs of our main results.

A.1 Unstructured multinomial

In this section we consider the set of matrices $\Theta = \{X \in \mathbb{R}^{k\times d} : \sum_{i,j} X_{i,j} = 1, X_{i,j} \geq 0\}$. We will first show Le Cam's method can not obtain the optimal minimax lower bound.

Proposition 14. *Consider the set, $\Theta = \{X \in \mathbb{R}^{k\times d} : \sum_{i,j} X_{i,j} = 1, X_{i,j} \geq 0\}$. Then the minimax risk satisfies*

$$\min_{\hat{X}} \max_{X\in\Psi} \mathbf{E}\, \|\hat{X} - X\|_{\mathrm{TV}} \geq C/\sqrt{n},$$

for a constant $C \geq 1$. Furthermore, the $\Omega(\sqrt{1/n})$ lower bound cannot be improved with Le Cam's method.

Proof of Proposition 14. We will use the framework of Le Cam's two-point argument [LC12]. For any pair of distributions $X^{(1)}, X^{(2)} \in \Psi$,

$$\min_{\hat{X}} \max_{X\in\Psi} \mathbf{E}\, \|\hat{X} - X\|_{\mathrm{TV}} \geq \frac{1}{4}\|X^{(1)} - X^{(2)}\|_{\mathrm{TV}} \exp\left(-nD_{\mathrm{KL}}(X^{(1)} \parallel X^{(2)})\right). \tag{9}$$

We then find the maximal lower bound of R^* can be reformulated as an optimization over $X^{(1)}$ and $X^{(2)}$ given as:

$$\max\left\{\|X^{(1)} - X^{(2)}\|_{\mathrm{TV}} : D_{\mathrm{KL}}(X^{(1)} \parallel X^{(2)}) \leq 1/n\right\}. \tag{10}$$

Pinsker's inequality states that $2\|X^{(1)} - X^{(2)}\|_{\mathrm{TV}}^2 \leq D_{\mathrm{KL}}(X^{(1)} \parallel X^{(2)})$ (see Equation (2)). This indicates that an upper bound for Equation (10) can be given as

$$\max\left\{\sqrt{\frac{1}{2}D_{\mathrm{KL}}(X^{(1)} \parallel X^{(2)})} : D_{\mathrm{KL}}(X^{(1)} \parallel X^{(2)}) \leq 1/n\right\}.$$

We then find that the optimal lower bound of the minimax error is of the form $R^* \gtrsim 1/\sqrt{n}$. All that remains is to construct $X^{(1)}, X^{(2)}$ to achieve error rate $C/\sqrt{n}$. We set $X^{(1)} = [1/2 - \sqrt{1/n}, 1/2 + \sqrt{1/n}, 0, \ldots, 0]$ and $X^{(2)} = [1/2 + \sqrt{1/n}, 1/2 - \sqrt{1/n}, 0, \ldots, 0]$. We then find that $\|X^{(1)} - X^{(2)}\|_{\mathrm{TV}} = 2/\sqrt{n}$. To upper bound the KL-divergence, we first note that from Taylor series of $\log(\frac{1+x}{1-x}) \leq Cx$ where $C < 1$. Then a simple calculation yields $D_{\mathrm{KL}}(X^{(1)} \parallel X^{(2)}) = O(1/n)$. Plugging in our calculations for the total variation lower bound and KL divergence upper bound into Equation (9) completes the proof. ■

To show that the $\Omega(1/\sqrt{n})$ lower bound from Proposition 14 is not optimal, we will use Fano's inequality to obtain an asymptotic improvement.

Theorem 15. *Consider the set $\Theta = \{X \in \mathbb{R}^{k\times d} : \sum_{i,j} X_{i,j} = 1, X_{i,j} \geq 0\}$. Then the minimax risk satisfies*

$$\min_{\hat{X}} \max_{X\in\Psi} \|\hat{X} - X\|_{\mathrm{TV}} \geq C\sqrt{dk/n},$$

for a constant $C \leq 0.35$ with probability at least $1/2$.

Proof of Theorem 15. We construct $\{X^{(1)}, \ldots, X^{(M)}\} \subset \mathbb{R}^d$ where each entry of $X^{(k)}$ takes values in $\{\mathrm{e}^{-\varepsilon}, \mathrm{e}^{\varepsilon}\}$ and $\|X^{(k)}\|_0 = d$ for all $1 \leq k \leq d$. We now show that $\|X^{(u)} - X^{(v)}\|_{\mathrm{TV}} \geq \Omega(\varepsilon)$ for all $u \neq v$. From our construction,

$$\|X^{(u)} - X^{(v)}\|_{\mathrm{TV}} = \frac{\exp(\varepsilon) - \exp(-\varepsilon)}{\frac{d}{2}\left(\exp(\varepsilon) + \exp(-\varepsilon)\right)} \cdot \|X^{(u)} - X^{(v)}\|_0 = \frac{2\tanh(\varepsilon)}{d} \cdot \|X^{(u)} - X^{(v)}\|_0 \geq \frac{\varepsilon}{d} \cdot \|X^{(u)} - X^{(v)}\|_0. \tag{11}$$

We will now show there exists a subset of all permutations such that $\|X^{(u)} - X^{(v)}\|_0 = \Omega(d)$ and $\log M \geq \Omega(d)$ with results from coding theory. We set $\|X^{(u)} - X^{(v)}\|_0 = cdk$ where $c = O(1)$ is a small constant.

From the Gilbert-Varshamov(GV) bound in Lemma 27, there exists a subset of size $A(d, cd)$ such that $\|X^{(u)} - X^{(v)}\|_0 \geq cd$ for all $i \neq j$, which implies $\|X^{(u)} - X^{(v)}\|_{\mathrm{TV}} \geq c\varepsilon$ from Equation (11). From the lemmata, $A(d, cd) \geq 2^{d(1-h_2(c))}$. We thus find $\log M \geq \Omega(d)$. Then from Equations (5) and (6),

$$\mathbf{Pr}\left(\min_{\hat{X}} \max_{X \in \Psi} \|X - \hat{X}\|_{\mathrm{TV}} \geq c\varepsilon\right) \geq 1 - \frac{\mathcal{I}(B, Y) + \log 2}{2 \log M}.$$

We note that the mutual information can be bounded from above as $\mathcal{I}(B; Y) \leq n \cdot \max_{u,v} D_{\mathrm{KL}}(X^{(u)} \parallel X^{(v)})$ [SC21]. We now upper bound the Fischer information. Let $\widetilde{X} = X / \sum_{i,j} X_{i,j}$ denote the multinomial distribution parameterized by the softmax of X so that for any outcome Y, we have $\widetilde{X}(Y) = (f(X))(Y)$. From the definition of the KL-divergence, we have for any $1 \leq u, v \leq M$ s.t. $u \neq v$,

$$\begin{aligned} D_{\mathrm{KL}}(\widetilde{X}^{(u)} \parallel \widetilde{X}^{(v)}) &= \sum\nolimits_{Y} \widetilde{X}^{(u)}(Y) \log\left(\widetilde{X}^{(u)}(Y) / \widetilde{X}^{(v)}(Y)\right) \\ &= \sum\nolimits_{Y} \widetilde{X}^{(u)}(Y) \sum\nolimits_{i,j} Y_{i,j} \left(X^{(u)}_{i,j} - X^{(v)}_{i,j} + \log\left(\sum\nolimits_{i,j} X^{(v)}_{i,j}\right) - \log\left(\sum\nolimits_{i,j} X^{(u)}_{i,j}\right)\right), \end{aligned}$$

where the second equality follows from Equation (3). We now note that finite sums can be interchanged, and obtain

$$D_{\mathrm{KL}}(\widetilde{X}^{(u)} \parallel \widetilde{X}^{(v)}) = \sum\nolimits_{i,j} f(X^{(u)})_{i,j} \left(X^{(u)}_{i,j} - X^{(v)}_{i,j} + \log\left(\sum\nolimits_{i,j} X^{(v)}_{i,j}\right) - \log\left(\sum\nolimits_{i,j} X^{(u)}_{i,j}\right)\right).$$

We note that since $X^{(v)}$ is a permutation of $X^{(u)}$, it follows that $\sum_{i,j} X^{(u)}_{i,j} = \sum_{i,j} X^{(v)}_{i,j}$. Define $\Delta_{i,j} \coloneqq X^{(u)}_{i,j} - X^{(v)}_{i,j}$. The KL divergence simplifies to

$$D_{\mathrm{KL}}(\widetilde{X}^{(u)} \parallel \widetilde{X}^{(v)}) = \sum\nolimits_{i,j} f(X^{(u)})_{i,j} \cdot \Delta_{i,j}.$$

The sum is maximized when the Hamming distance between $X^{(u)}$ and $X^{(v)}$. When the Hamming distance is maximized, then $\Delta_{i,j} = \varepsilon - (-\varepsilon) = 2\varepsilon$ or $\Delta_{i,j} = (-\varepsilon) - \varepsilon = -2\varepsilon$.

$$\max_{u,v \leq M, u \neq v} D_{\mathrm{KL}}(\widetilde{X}^{(u)} \parallel \widetilde{X}^{(v)}) \leq d\left(\frac{2\varepsilon(\exp(\varepsilon) - \exp(-\varepsilon))}{d(\exp(\varepsilon) + \exp(-\varepsilon))}\right) = 2\varepsilon \tanh(\varepsilon) \leq 2\varepsilon^2,$$

where we use the fact that maximally all the elements are mismatched (Hamming distance of d). Combining these results, we have

$$\mathbf{Pr}\left(\min_{\hat{X}} \max_{X \in \Psi} \|X - \hat{X}\|_{\mathrm{TV}} \geq c\varepsilon\right) \geq 1 - \frac{2n\varepsilon^2 + \log 2}{(1 - h_2(c))kd \log 2}.$$

We then set $\varepsilon = O(\sqrt{kd/n})$ and obtain $\mathfrak{R}^* \geq \Omega(\sqrt{kd/n})$ with probability at least $1/2$. Our proof is complete. ■

Lemma 16. *Consider the set, $\Theta = \{X \in \mathbb{R}^{k \times d} : \sum_{i,j} X_{i,j} = 1, X_{i,j} \geq 0\}$. Suppose $\widetilde{X}(Y)$ is the maximum likelihood estimator. Then,*

$$\max_{X \in \Psi} \mathbf{E} \|\widetilde{X}(Y) - X\|_{\mathrm{TV}} \leq 2\sqrt{kd/n}.$$

Proof of Lemma 16. By the linearity of expectation,

$$\mathbf{E}\|\widetilde{X}(Y) - X\|_{\mathrm{TV}} = \frac{1}{2n} \sum\nolimits_{i,j} \mathbf{E}\,|Y_{i,j} - \mathbf{E}\,Y_{i,j}|.$$

Note that $\mathbf{Var}(Y_{i,j}) = nX_{i,j}(1 - X_{i,j})$, then from Chebyshev's inequality [Fel91],

$$\mathbf{E}\,|Y_{i,j} - \mathbf{E}[Y_{i,j}]| = \int_0^\infty \mathbf{Pr}\{|Y_{i,j} - \mathbf{E}[Y_{i,j}]| \geq t\}\mathrm{d}t \leq C + \int_C^\infty \frac{nX_{i,j}(1 - X_{i,j})}{t^2}\mathrm{d}t = C + \frac{nX_{i,j}(1 - X_{i,j})}{C}.$$

Optimizing over C, we set $C = \sqrt{nX_{i,j}(1-X_{i,j})}$,

$$\frac{1}{2n}\sum\nolimits_{i,j}\mathbf{E}\left|Y_{i,j}-\mathbf{E}[Y_{i,j}]\right| \leq \frac{1}{n}\sum\nolimits_{i,j}\sqrt{nX_{i,j}(1-X_{i,j})}.$$

We then find the maximizing $X \in \Theta$ as $X = (1/kd, \ldots, 1/kd)$. We then obtain

$$\max_{X\in\Psi}\mathbf{E}\|\hat{X}-X\|_{\mathrm{TV}} \leq \sqrt{kd/n}.$$

Our proof is complete. ■

Theorem 17. *Let $\Theta = \{X \in \mathbb{R}^{k\times d} : \sum_{i,j} X_{i,j} = 1, X_{i,j} \geq 0\}$. Suppose we observe n i.i.d one-hot samples $Y_1, \ldots, Y_n$. Then,*

$$\min_{\hat{X}}\max_{X\in\Psi}\|\hat{X}-X\|_{\mathrm{TV}} \leq 2\sqrt{\frac{kd}{n}} + \frac{\sqrt{2\log(1/\delta)}}{n},$$

with probability at least $1-\delta$.

Proof of Theorem 17. Suppose $\widetilde{X}(Y_1, \ldots, Y_n) = \sum_k X_k/n$ represents the empirical estimator. From standard arguments, we have

$$\min_{\hat{X}}\max_{X\in\Psi}\mathbf{E}\,\|\hat{X}-X\|_{\mathrm{TV}} \leq \max_{X\in\Psi}\mathbf{E}\,\|\widetilde{X}-X\|_{\mathrm{TV}}$$

We define the function :

$$f(Y_1, \ldots, Y_n) = \|\widetilde{X}(Y_1, \ldots, Y_n) - X\|_{\mathrm{TV}}.$$

Suppose Δ represents the canonical basis of $\mathbb{R}^{d\times k}$. For each $1 \leq k \leq n$,

$$\sup_{Y_k'\in\Delta}|f(Y_1, \ldots, Y_{k-1}, Y_k, Y_{k+1}, \ldots, Y_n) - f(Y_1, \ldots, Y_{k-1}, Y_k', Y_{k+1}, \ldots, Y_n)| \leq 2/n.$$

Using the bounded differences with Lemma 28, we obtain

$$\mathbf{Pr}\left(f(Y_1, \ldots, Y_n) - \mathbf{E}[f(Y_1, \ldots, Y_n)] \geq \varepsilon\right) \leq \exp(-\varepsilon^2 n^2/2).$$

From Lemma 16, we have $\mathbf{E}\,\|\widetilde{X}-X\|_{\mathrm{TV}} \leq 2\sqrt{kd/n}$ for all $X \in \Psi$, we then obtain

$$\|\widetilde{X}-X\|_{\mathrm{TV}} \leq 2\sqrt{\frac{kd}{n}} + \frac{\sqrt{2\log(1/\delta)}}{n},$$

with probability at least $1-\delta$. Our proof is complete. ■

A.2 Low-rank probability matrix

Theorem 18. *Consider the set, Θ, consisting of multinomial distributions parameterized with matrices with rank less than or equal to r. Then the minimax risk satisfies,*

$$\min_{\hat{X}}\max_{X\in\Theta}\|\widehat{X}(Y)-X\|_{\mathrm{TV}} \geq C\sqrt{r(d+k-r)/n},$$

with probability at least $1/2$, where $C \leq c(1-h_2(c))/4 \leq 0.126$.

Proof of Theorem 18. We first construct the packing set of Θ. We set $m \coloneqq (r/2)(k+d-(r/2))$. We choose a binary code $\mathcal{V} \subset \{0,1\}^m$ with blocklength m and minimum Hamming distance at least cm for a constant $c < 1/2$ and size $M \geq \exp((1-h_2(c))m)$ guaranteed by the GV bound (cf. Lemma 27). elements $\{X^{(1)}, \ldots, X^{(M)}\}$ such that $\mathrm{rank}(X^{(u)}) \leq r$ for all $u \leq M$ and $\|X^{(u)} - X^{(v)}\|_{\mathrm{TV}} \geq \Omega(\delta)$ for all $u, v \leq M$ and $u \neq v$.

$$\mathbf{Pr}\left(\min_{\hat{X}}\max_{X\in\Theta}\|\widehat{X}(Y)-X\|_{\mathrm{TV}} \geq \delta/2\right) \geq \min_{\tilde{X}}\mathbf{Pr}[\tilde{X} \neq B],$$

where B is uniformly distributed over $\{X^{(1)}, \dots, X^{(M)}\}$. Fano's inequality [YB99, Yu97] gives us,

$$\min_{\tilde{X}} \mathbf{Pr}[\tilde{X} \neq B] \geq 1 - \frac{\mathcal{I}(Y;B) + \log 2}{\log M}. \tag{12}$$

Let $m \coloneqq r(k+d-r/2)/2$. Consider $S = \{1, \dots, m\}$. Then let $\{U_1, \dots, U_M\}$ represent a subset of the $\binom{m}{m/2}$ size $m/2$ combinations of S. Then for each U_i, let

$$\tilde{X}^{(v)}_{i,j} = \begin{cases} \mathrm{e}^{\delta}, & (i,j) \in U^{(v)}, (i \leq r/2 \text{ or } j \leq r/2), \\ \mathrm{e}^{-\delta}, & (i,j) \notin U^{(v)}, (i \leq r/2 \text{ or } j \leq r/2), \\ 0, & i > r/2 \text{ and } j > r/2, \end{cases}$$

and normalize $X^{(v)} = \tilde{X}^{(v)} / \sum_{i,j} \tilde{X}^{(v)}_{i,j}$ to a probability matrix. We can write $X^{(v)}$ in block notation as

$$X^{(v)} = \begin{bmatrix} A & B \\ C & 0 \end{bmatrix}, \; A \in \mathbb{R}^{(r/2)\times(r/2)}, \; B \in \mathbb{R}^{(r/2)\times(d-r/2)}, \; C \in \mathbb{R}^{(k-r/2)\times(r/2)}$$

We can upper bound the rank of $X^{(v)}$ using the block partition,

$$\operatorname{rank}(X^{(v)}) \leq \operatorname{rank}([A,B]) + \operatorname{rank}([C,0]) \leq r/2 + r/2 = r.$$

Therefore $X^{(v)} \in \Theta$ for all v. We now consider a lower bound for the separation, consider any distinct $u, v \leq M$,

$$\|X^{(u)} - X^{(v)}\|_{\mathrm{TV}} \gtrsim m \tanh(\delta) \geq m\delta.$$

In the above we use the fact that the Hamming distance between $X^{(u)}$ and $X^{(v)}$ is $\Omega(m)$. We now upper bound the mutual information. For any distinct $u, v \leq M$,

$$\begin{aligned} D_{\mathrm{KL}}(X^{(u)} \parallel X^{(v)}) &= \sum\nolimits_Y X^{(u)}(Y) \log\left(X^{(u)}_{i,j}(Y) / X^{(v)}_{i,j}(Y)\right) \\ &= \sum\nolimits_Y X^{(u)}(Y) \sum\nolimits_{i,j} Y_{i,j} \left(\tilde{X}^{(u)}_{i,j} - \tilde{X}^{(v)}_{i,j} + \log\left(\sum\nolimits_{i,j} \tilde{X}^{(u)}_{i,j}\right) - \log\left(\sum\nolimits_{i,j} \tilde{X}^{(v)}\right)\right). \end{aligned}$$

By the construction of the packing set, $X^{(u)}$ and $X^{(v)}$ are just permutations of each other. We can interchange the sums, then use the law of the unconscious statistician to obtain

$$D_{\mathrm{KL}}(X^{(u)} \parallel X^{(v)}) = \sum\nolimits_{i,j} X^{(u)}_{i,j} \left(\tilde{X}^{(u)}_{i,j} - \tilde{X}^{(v)}_{i,j}\right) \coloneqq \sum\nolimits_{i,j} X^{(u)}_{i,j} \Delta_{i,j}.$$

For each i, j, we find that $\Delta_{i,j}$ takes values in $\{0, -2\delta, 2\delta\}$, and $\Delta_{i,j} = 2\delta$ only if $X^{(u)}_{i,j} = \mathrm{e}^{\delta}$ and $\Delta_{i,j} = -2\delta$ only if $X^{(u)}_{i,j} = \mathrm{e}^{-\delta}$. We note that the Hamming distance between $X^{(u)}_{i,j}$ and $X^{(v)}_{i,j}$ is $O(m)$,

$$D_{\mathrm{KL}}(X^{(u)} \parallel X^{(v)}) \lesssim \delta \tanh(\delta) \leq 2\delta^2. \tag{13}$$

We can then note from standard arguments that $\mathcal{I}(B;Y) \leq n \cdot \max_{i,j} D_{\mathrm{KL}}(X^{(i)} \parallel X^{(j)})$ and from Equation (13) if follows that $I(B;Y) \leq 2n\delta^2$. From the GV-bound in Lemma 27, we find that there exists a subset such that $\log M \geq (1 - h_2(c))r(d+k-r)$ and $\|X^{(i)} - X^{(j)}\|_{\mathrm{TV}} \geq c\delta$ for a constant c. Plugging this in to Equation (12),

$$\min_{\tilde{X}} \mathbf{Pr}[\tilde{X} \neq B] \geq 1 - \frac{2n\delta^2 + \log 2}{(1 - h_2(c))r(d+k)} \geq \frac{1}{2},$$

when we set $\delta \leq \sqrt{Cr(d+k)/n}$ where $C \leq (1 - h_2(c))/4c$. Our proof is complete ∎

Theorem 19. *Consider the set of matrices* $\Theta = \left\{X \in \mathbb{R}^{k\times d} : \sum_{i,j} X_{i,j} = 1, X_{i,j} \geq 0, \operatorname{rank}(X) \leq r\right\}$. *Suppose* $Y = \sum_i Y_i/n$ *where* $Y_i \in \mathbb{R}^{k\times d}$ *is a one-hot matrix sampled from* X. *Then the minimax risk satisfies,*

$$\min_{\hat{X}} \max_{X\in\Theta} \|\widehat{X}(Y) - X\|_{\mathrm{TV}} \leq 4\sqrt{r(k+d)\log(k+d)/n},$$

with probability at least $1 - 2/(k+d)$ *when* $n \geq 4\nu(k+d)\log(k+d)$.

Our proof is by using the nuclear norm penalized maximum log-likelihood estimator introduced in [LKMS14].

Proof of Theorem 19. We define our estimator as the minimizer to the following minimization problem

$$\widehat{X} = \arg\min_{M \in \Theta} \left\{ -\Phi_Y(M) + \lambda \|M\|_{\sigma,1} \right\}, \tag{14}$$

where the loss is defined as

$$\Phi_Y(M) = \sum\nolimits_{i,j} Y_{i,j} \log M_{i,j}.$$

By the optimality of $\widehat{X}$, we can note that

$$-\Phi_Y(\widehat{X}) + \lambda \|\widehat{X}\|_{\sigma,1} \leq -\Phi_Y(X) + \lambda \|X\|_{\sigma,1},$$

and therefore

$$\Phi_Y(X) - \Phi_Y(\widehat{X}) \leq \lambda \left(\|X\|_{\sigma,1} - \|\widehat{X}\|_{\sigma,1} \right). \tag{15}$$

We can use Lemma 29 to upper bound the difference in nuclear norms,

$$\|X\|_{\sigma,1} - \|\widehat{X}\|_{\sigma,1} \leq \|\mathcal{P}_X(\widehat{X} - X)\|_{\sigma,1}.$$

We note that $\operatorname{rank}(\mathcal{P}_X(\widehat{X} - X)) \leq \operatorname{rank}(X)$, therefore

$$\|\mathcal{P}_X(\widehat{X} - X)\|_{\sigma,1} \leq \sqrt{\operatorname{rank}(X)} \|\widehat{X} - X\|_{\sigma,2},$$

where we use the inequalities, $\|A\|_{\sigma,1} \leq \sqrt{\operatorname{rank}(A)} \|A\|_{\sigma,2}$ and $\|AB\|_{\sigma,\infty} \leq \|A\|_{\sigma,\infty} \|B\|_{\sigma,\infty}$. We can now upper bound the KL-divergence between X and $\widehat{X}$,

$$D_{\mathrm{KL}}(X \parallel \widehat{X}) = \mathbf{E}\left[\Phi_Y(X) - \Phi_Y(\widehat{X}) \right] \leq 4\lambda \sqrt{\operatorname{rank}(X)} \|X - \widehat{X}\|_{\sigma,2},$$

where in the second inequality we use Markov's inequality with failure probability $1/4$. Then from Pinsker's inequality (see Equation (2)),

$$\|X - \widehat{X}\|_{\sigma,2} \leq \sqrt{2 D_{\mathrm{KL}}(X \parallel \widehat{X})}.$$

Combining the inequalities, we obtain

$$D_{\mathrm{KL}}(X \parallel \widehat{X}) \leq 4\lambda \sqrt{\operatorname{rank}(X) D_{\mathrm{KL}}(X \parallel \widehat{X})}.$$

Rearranging,

$$\sqrt{D_{\mathrm{KL}}(X \parallel \widehat{X})} \leq 4\lambda \sqrt{\operatorname{rank}(X)}.$$

We then apply Pinsker's inequality (see Equation (2)) once more,

$$\|X - \widehat{X}\|_{\mathrm{TV}} \leq \lambda \sqrt{32 \operatorname{rank}(X)}.$$

Then, using our choice of λ in Lemma 25, we obtain

$$\|X - \widehat{X}\|_{\mathrm{TV}} \leq \sqrt{16 r (k+d) \log(k+d)/n}.$$

Our proof is complete. ■

Lemma 20. *Suppose the regularization parameter for the minimization procedure in Theorem 19 is chosen as*

$$\lambda \geq \sqrt{\frac{4\nu(k+d)\log(k+d)}{n}},$$

then KKT conditions are satisfied for X with probability at least $1 - \delta$.

Proof of Lemma 20. The Lagrangian function [BSS06] for the minimization problem defined in Equation (14) is given as

$$\mathcal{L}(X, \xi, \Lambda) = -\Phi_Y(X) + \lambda \|X\|_{\sigma,1} + \xi \left(\sum\nolimits_{i,j} X_{i,j} - 1 \right) - \sum\nolimits_{i,j} \Lambda_{i,j} X_{i,j}.$$

We now list Karush-Kuhn-Tucker (KKT) conditions [Kar39, KT13] for X to be an optimal point in the optimization routine. We first give the subdifferential for the nuclear norm [HTW15] of a matrix X with singular value decomposition $U\Sigma V^\top$,

$$\partial \|X\|_{\sigma,1} = \left\{ G \in \mathbb{R}^{k \times d} \mid G = UV^\top + W, \text{ where } U^\top W = 0,\ WV = 0,\ \|W\|_{\sigma,\infty} \leq 1 \right\}. \tag{16}$$

The dual feasability condition is

$$\Lambda_{i,j} \geq 0, \text{ for } 1 \leq i \leq k,\ 1 \leq j \leq d. \tag{17}$$

The complementary slackness condition is

$$-\sum\nolimits_{i,j} \Lambda_{i,j} X_{i,j} = 0. \tag{18}$$

Stationarity is satisfied when,

$$-\nabla \Phi_Y(X) + \lambda G + \xi \mathbf{1}\mathbf{1}^\top - \Lambda = 0,$$

for a $G \in \partial \|X\|_{\sigma,1}$. This implies at each i, j,

$$-Y_{i,j}/X_{i,j} + \lambda G_{i,j} + \xi - \Lambda_{i,j} = 0.$$

Rearranging, we obtain

$$-Y_{i,j} + \lambda G_{i,j} X_{i,j} + \xi X_{i,j} - \Lambda_{i,j} X_{i,j} = 0.$$

From dual feasability (see Equation (17)) and complementary slackness (see Equation (18)), it follows that $\Lambda_{i,j} X_{i,j} = 0$ for all $1 \leq i \leq k$, $1 \leq j \leq d$. Therefore,

$$\lambda G \circ X = Y - \xi X.$$

Taking the spectral norm of both sides,

$$\lambda \|G \circ X\|_{\sigma,\infty} = \|Y - \xi X\|_{\sigma,\infty}.$$

We note that from [HJ94, Theorem 5.5.1] for any two matrices $A, B \in \mathbb{R}^{k \times d}$, $\|A \circ B\|_{\sigma,\infty} \leq \|A\|_{\sigma,\infty} \|B\|_{\sigma,\infty}$, this gives us the inequality

$$\lambda \|G\|_{\sigma,\infty} \|X\|_{\sigma,\infty} \geq \|Y - \xi X\|_{\sigma,\infty}.$$

We then use the properties of G in Equation (16) and note that $\min_X \|X\|_{\sigma,\infty} \geq 1/\sqrt{kd}$,

$$2\lambda \geq \sqrt{kd} \|Y - \xi X\|_{\sigma,\infty}.$$

Taking $\xi \coloneqq 1$, we use Lemma 26 and obtain,

$$\lambda \geq \sqrt{\frac{4\nu(k+d)\log((k+d)/\delta)}{n}},$$

with probability at least $1 - \delta$ when $n \geq 4\nu(k+d)\log(k+d)$. Our proof is complete. ■

A.3 Low-rank underlying softmax

We first derive a lower bound on the minimax risk with Fano's inequality.

Theorem 21. *Consider the set, $\Theta = \{X \in \mathbb{R}^{k \times d} : \operatorname{rank}(X) \leq r\}$ and f represents the softmax function over all the matrix elements. Then the minimax risk satisfies,*

$$\min_{\widehat{X}} \max_{X \in \Theta} \|f(\widehat{X}) - f(X)\|_{\mathrm{TV}} \geq \sqrt{C(r-1)(d+k-(r-1)/2)/n},$$

with probability exceeding $1/2$.

Proof follows the same steps as Theorem 18 with a modified packing set.

Proof of Theorem 21. Consider a packing set $\{f(X^{(1)}), \dots, f(X^{(M)})\}$ such that $\operatorname{rank}(X^{(v)}) \le r$ for all $v \le M$ and $\|f(X^{(u)}) - f(X^{(v)})\|_{\mathrm{TV}} \ge \Omega(\delta)$ for all $1 \le u, v \le M$ and $u \ne v$,

$$\mathbf{Pr}\left(\min_{\hat{X}} \max_{X \in \Theta} \|f(\hat{X}) - f(X)\|_{\mathrm{TV}} \ge \delta/2\right) \ge \min_{\hat{X}} \mathbf{Pr}[f(\tilde{X}) \ne B],$$

where B is uniformly distributed over $\{f(X^{(1)}), \dots, f(X^{(M)})\}$. Then from Fano's inequality [YB99, Yu97],

$$\min_{\tilde{X}} \mathbf{Pr}[\tilde{X} \ne B] \ge 1 - \frac{\mathcal{I}(Y; B) + \log 2}{\log M}. \tag{19}$$

Let $m = r(d + k - r/2)/2$. Consider $S = \{1, \dots, m\}$. Then let $\{U_1, \dots, U_M\}$ represent a subset of the $\binom{m}{m/2}$ size m combinations of S. Then for each U_i, let

$$X_{i,j}^{(v)} = \begin{cases} \varepsilon \cdot \left(2 \cdot \mathbf{1}\{(i,j) \in U^{(v)}\} - 1\right) & 1 \le i \le (r-1)/2 \text{ or } 1 \le j \le (r-1)/2 \\ -\xi & i > (r-1)/2 \text{ and } j > (r-1)/2 \end{cases}.$$

We can partition $f(X^{(v)})$ as a block matrix

$$f(X^{(v)}) = \begin{bmatrix} A & B \\ C & -\xi \cdot I \end{bmatrix}.$$

We verify that each matrix in the packing set has rank at most r with the Schur complement,

$$\operatorname{rank}(X^{(v)}) = \operatorname{rank}(A) + \operatorname{rank}(-\xi \cdot I - CA^{-1}B) \le (r-1)/2 + 1 + (r-1)/2 = r.$$

We first verify the packing distance lower bound. Set $m \coloneqq (r-1)(k + d + (r-1)/2)/2$,

$$\|f(X^{(i)}) - f(X^{(j)})\|_{\mathrm{TV}} \ge \frac{\mathrm{e}^{\varepsilon} - \mathrm{e}^{-\varepsilon}}{m(\mathrm{e}^{\varepsilon} + \mathrm{e}^{-\varepsilon}) + (kd - m) \cdot \mathrm{e}^{-\xi}} \cdot \|f(X^{(i)}) - f(X^{(j)})\|_0,$$

then when we set $\xi \ge \log(\varepsilon + ((kd - m)/m))$, we find

$$\|f(X^{(i)}) - f(X^{(j)}\|_{\mathrm{TV}} \ge \varepsilon \cdot \|f(X^{(i)}) - f(X^{(j)}\|_0.$$

We now upper bound the Fischer information. Define $\Delta_{i,j} = X_{i,j}^{(k_1)} - X_{i,j}^{(k_2)}$. The KL divergence is then

$$D_{\mathrm{KL}}(f(X^{(k_1)}) \parallel f(X^{(k_2)})) = \sum\nolimits_{i,j} f(X_{i,j}^{(k_1)}) \cdot \Delta_{i,j}.$$

From the construction, if $\Delta_{i,j}$ is non-zero, then $\Delta_{i,j} = 2\varepsilon / \sum_{i,j} X_{i,j}^{(k_1)}$ or $\Delta_{i,j} = -2\varepsilon / \sum_{i,j} X_{i,j}^{(k_1)}$. Set $m \coloneqq r(k + d + r/2)/2$,

$$\max_{i \ne j} D_{\mathrm{KL}}(X^{(k_1)} \parallel X^{(k_2)}) \le \frac{2m\varepsilon(\mathrm{e}^{\varepsilon} - \mathrm{e}^{-\varepsilon})}{m(\mathrm{e}^{\varepsilon} + \mathrm{e}^{-\varepsilon}) + (kd - m) \cdot \mathrm{e}^{-\xi}}.$$

If $\xi \ge \log(\varepsilon + ((kd - m)/m))$, then

$$\max_{u \ne v} D_{\mathrm{KL}}(X^{(u)} \parallel X^{(v)}) \le 2\varepsilon^2.$$

The GV bound (see Lemma 27) guarantees the existence of a subset such that $\log M \ge (1 - h_2(c))m$, and $\|X^{(u)} - X^{(v)}\|_{\mathrm{TV}} \ge c\varepsilon$, for all $u \ne v$. We use our bounds for the Fischer information and packing set cardinality in Equation (19) and obtain

$$\mathbf{Pr}\left(\min_{\hat{X}} \max_{X \in \Theta} \|f(\hat{X}) - f(X)\|_{\mathrm{TV}} \ge c\varepsilon/2\right) \ge 1 - \frac{2n\varepsilon^2 + \log 2}{m \log 2}.$$

We then set $\varepsilon = \sqrt{m/n}$ to complete the proof. . ■

Theorem 22. *Consider the set of matrices*

$$\Gamma = \left\{ X \in \mathbb{R}^{k \times d} : \sum\nolimits_{i,j} X_{i,j} = 0, |X_{i,j}| \leq C, \operatorname{rank}(X) \leq r \right\}.$$

Suppose f is the softmax function. Then the minimax risk satisfies,

$$\min_{\widehat{X}} \max_{X \in \Gamma} \|f(\widehat{X}) - f(X)\|_{\mathrm{TV}} \leq 16c\sqrt{r(k+d)\log(k+d)/n},$$

with probability at least $1 - 1/8 - 2/(k+d)$ when $n \geq 4\nu(k+d)\log(k+d)$.

Our proof is by using the nuclear norm penalized maximum log-likelihood estimator introduced in [LKMS14].

Proof of Theorem 22. Consider the set of matrices,

$$G = \left\{ M \in \mathbb{R}^{k \times d} : \max\nolimits_{i,j} |M_{i,j}| \leq C, \sum\nolimits_{i,j} M_{i,j} = 0 \right\}.$$

We define our estimator as the minimizer to the following minimization problem

$$\widehat{X} = \arg\min_{M \in G} \left\{ -\Phi_Y(f(M)) + \lambda \|M\|_{\sigma,1} \right\}, \tag{20}$$

where the loss is defined as

$$\Phi_Y(f(M)) = \sum\nolimits_{i,j} Y_{i,j} \log(f(M))_{i,j}.$$

By the optimality of $\widehat{X}$, we can note that

$$-\Phi_Y(f(\widehat{X})) + \lambda \|\widehat{X}\|_{\sigma,1} \leq -\Phi_Y(f(X)) + \lambda \|X\|_{\sigma,1},$$

and therefore

$$\Phi_Y(f(X)) - \Phi_Y(f(\widehat{X})) \leq \lambda \left(\|X\|_{\sigma,1} - \|\widehat{X}\|_{\sigma,1} \right). \tag{21}$$

We can use Lemma 29 to upper bound the difference in nuclear norms,

$$\|X\|_{\sigma,1} - \|\widehat{X}\|_{\sigma,1} \leq \|\mathcal{P}_X(\widehat{X} - X)\|_{\sigma,1}.$$

We note that $\operatorname{rank}(\mathcal{P}_X(\widehat{X} - X)) \leq \operatorname{rank}(X)$, therefore

$$\|\mathcal{P}_X(\widehat{X} - X)\|_{\sigma,1} \leq \sqrt{\operatorname{rank}(X)} \|\widehat{X} - X\|_{\sigma,2},$$

where we use the inequalities, $\|A\|_{\sigma,1} \leq \sqrt{\operatorname{rank}(A)} \|A\|_{\sigma,2}$ and $\|AB\|_{\sigma,\infty} \leq \|A\|_{\sigma,\infty} \|B\|_{\sigma,\infty}$. Consider any M such that Equation (21) holds, then

$$D_{\mathrm{KL}}(f(X) \parallel f(M)) = \mathbf{E}\left[\Phi_Y(f(X)) - \Phi_Y(f(M))\right] \leq 8\lambda\sqrt{\operatorname{rank}(X)} \|X - M\|_{\sigma,2},$$

where in the second inequality we use Equation (21) and Markov's inequality with failure probability at most $1/8$. We can then note that Equation (21) holds for $\widehat{X}$. Then from Lemma 23,

$$\|X - \widehat{X}\|_{\sigma,2} \leq c\sqrt{128kdD_{\mathrm{KL}}(f(X) \parallel f(\widehat{X}))}.$$

Combining the inequalities, we obtain

$$D_{\mathrm{KL}}(f(X) \parallel f(\widehat{X})) \leq c\lambda\sqrt{128kd \operatorname{rank}(X) D_{\mathrm{KL}}(f(X) \parallel f(\widehat{X}))}.$$

Rearranging,

$$\sqrt{D_{\mathrm{KL}}(f(X) \parallel f(\widehat{X}))} \leq c\lambda\sqrt{128kd \operatorname{rank}(X)}.$$

We then apply Pinsker's inequality (see Equation (2)),

$$\|f(X) - f(\widehat{X})\|_{\mathrm{TV}} \leq 16c\lambda\sqrt{kd \operatorname{rank}(X)}.$$

Then, using our choice of λ in Lemma 25, we obtain

$$\|f(X) - f(\widehat{X})\|_{\mathrm{TV}} \leq 16c\sqrt{r(k+d)\log(k+d)/n}.$$

Our proof is complete. ∎

Lemma 23. *Define $G = \{X \in \mathbb{R}^{k\times d} : \sum_{i,j} X_{i,j} = 0\}$ such that Assumption 7 holds. Consider $X, M \in G$. Suppose f is the softmax function, then*

$$\|X - M\|_{\sigma,2} \leq \sqrt{2kd\mu^{-1} D_{\mathrm{KL}}(f(X) \parallel f(M))}.$$

Proof of Lemma 23. By the definition of KL divergence,

$$D_{\mathrm{KL}}(f(X) \parallel f(M)) = \sum\nolimits_{i,j} f(X)_{i,j} \log(f(X)_{i,j}/f(M)_{i,j}) = \mathrm{lse}(M) - \mathrm{lse}(X) - \langle \nabla \mathrm{lse}(X), M - X\rangle.$$

The integral form of Taylor remainder [Fir60] in the first order for a function $g : \mathbb{R} \mapsto \mathbb{R}$ that is at least twice differentiable is given as

$$g(1) - g(0) - g'(0) = \int_0^1 (1-t) g''(t) \mathrm{d}t \tag{22}$$

Let $\Delta \coloneqq M - X$ and define $g(t) \coloneqq \mathrm{lse}(X + t\Delta)$, then from Equation (22),

$$D_{\mathrm{KL}}(f(X) \parallel f(M)) = \int_0^1 (1-t)\Delta^\top J((1-t)X + tM)\Delta \,\mathrm{d}t. \tag{23}$$

We set $J((1-t)X + tM) = \mathrm{diag}(u) - uu^\top$, where $u \in \mathbb{R}^{kd}$ is a probability vector (i.e $\sum_{i,j} u_{i,j} = 1, u_{i,j} \geq 0$) representing the flattened softmax of $(1-t)X + tM$. Suppose $v \in \mathbb{R}^{kd}$ is a zero-sum vector. Then

$$v^\top(\mathrm{diag}(u) - uu^\top)v = \sum\nolimits_i u_i v_i^2 - \left(\sum\nolimits_i u_i v_i\right)^2.$$

We can expand the squared term as

$$\left(\sum\nolimits_i u_i v_i\right)^2 = \sum\nolimits_i u_i^2 v_i^2 + \sum\nolimits_{i\neq j} u_i u_j v_i v_j,$$

Rearranging, we obtain

$$v^\top(\mathrm{diag}(u) - uu^\top)v = \sum\nolimits_i v_i^2(u_i - u_i^2) - 2\sum\nolimits_{i<j} u_i u_j v_i v_j.$$

Expanding the first term,

$$\sum\nolimits_i v_i^2 u_i(1-u_i) = \sum\nolimits_i \sum\nolimits_{j\neq i} v_i^2 u_i u_j = \sum\nolimits_{i<j} (v_i^2 u_i u_j + v_j^2 u_i u_j).$$

We then find that

$$v^\top(\mathrm{diag}(u) - uu^\top)v = \sum\nolimits_{i<j} u_i u_j (v_i - v_j)^2 \geq \min_i u_i^2 \cdot \sum\nolimits_{i<j} (v_i - v_j)^2.$$

We can lower bound the summand with the fact that $\sum_i v_i = 0$.

$$2\sum\nolimits_{i<j}(v_i - v_j)^2 = \sum\nolimits_i \sum\nolimits_j (v_i - v_j)^2 = \sum\nolimits_{i,j} v_i^2 + \sum\nolimits_{i,j} v_j^2 - 2\sum\nolimits_{i,j} v_i v_j = 2kd\|v\|^2,$$

where in the final equality, we use the fact that $\sum_i v_i = 0$. We then obtain

$$v^\top(\mathrm{diag}(u) - uu^\top)v \geq kd\|v\|^2 \cdot \min_i u_i^2. \tag{24}$$

We then plug Equation (24) into Equation (23) and obtain

$$\int_0^1 (1-t)\Delta^\top J((1-t)X + tM)\Delta \,\mathrm{d}t \geq \frac{\mu}{kd}\int_0^1 (1-t)\|\Delta\|^2 \,\mathrm{d}t = \frac{\mu}{2kd}\|X - M\|_{\sigma,2}^2.$$

In the above, we use the convexity argument from Lemma 24. Rearranging, we find that

$$\|X - M\|_{\sigma,2} \leq \sqrt{2kd\mu^{-1} D_{\mathrm{KL}}(f(X) \parallel f(M))}.$$

Our proof is complete. ■

Lemma 24. *Consider the set* $\Theta = \{X \in \mathbb{R}^{k\times d} : \sum_{i,j} X_{i,j} = 0, f(X)_{u,v} \geq \mu/kd\}$ *such that* $f(X)_{u,v} \geq \mu/kd$ *for all* $u \leq k$ *and* $v \leq d$*. This set is convex.*

Proof of Lemma 24. Consider $X \in \Theta$ such that Assumption 7 holds. Then for all $u \leq k, v \leq d$,

$$X_{u,v} - \operatorname{lse}(X) \geq \log(\mu/kd). \tag{25}$$

Consider the convex combination $(1-t)A + tB$ for $A, B \in \Theta$,

$$(1-t)A_{u,v} + tB_{u,v} - \operatorname{lse}((1-t)A + tB) \geq (1-t)(A_{u,v} - \operatorname{lse}(A)) + t(B_{u,v} - \operatorname{lse}(B)) \geq \log(\mu/kd),$$

in the above, we use the convexity of the log-sum-exp function in the first inequality, and the second inequality follows from $A, B \in \Theta$, which implies that Equation (25) holds. ■

Lemma 25. *Suppose the regularization parameter for the minimization procedure in Theorem 22 is chosen as*

$$\lambda \geq \sqrt{\frac{4\nu \log((k+d)/\delta)}{n \min(k,d)}},$$

then KKT conditions are satisfied for X *with probability at least* $1 - 1/(k+d)$*.*

Proof of Lemma 25. The Lagrangian is given as

$$\mathcal{L}(X, \xi, \Lambda) = -\Phi(X) + \lambda \|X\|_{\sigma,1} + \xi \sum\nolimits_{i,j} X_{i,j} + \sum\nolimits_{i,j} \Lambda^+_{i,j}(X_{i,j} - C) + \sum\nolimits_{i,j} \Lambda^-_{i,j}(-X_{i,j} - C)$$

for a $G \in \partial \|X\|_{\sigma,1}$. The complementary slackness condition is

$$\Lambda^+_{i,j}(X_{i,j} - C) = 0, \quad \Lambda^-_{i,j}(-X_{i,j} - C) = 0.$$

The stationary condition is

$$-Y + f(X) + \lambda G + \gamma \mathbf{1}\mathbf{1}^\top = 0.$$

Consider the projection matrix Π that maps to the subspace orthogonal to $\{\mathbf{1}\}$, then applying the projection to both sides, we obtain

$$\Pi(-Y_{i,j} + f(X)_{i,j} + \lambda G_{i,j}) = 0.$$

Rearranging,

$$\lambda(\Pi G) = \Pi(Y - f(X)).$$

Taking the spectral norm of both sides,

$$\lambda \|\Pi G\|_{\sigma,\infty} = \|\Pi(Y - f(X))\|_{\sigma,\infty}.$$

We note that $\|G\|_{\sigma,\infty} \leq 2$ from Equation (16) and the projection operator is a contraction,

$$2\lambda \geq \|\Pi(Y - f(X))\|_{\sigma,\infty}.$$

We use Lemma 26 and obtain,

$$\lambda \geq \sqrt{\frac{4\nu \log((k+d)/\delta)}{n \min(k,d)}},$$

with probability at least $1 - \delta$ when $n \geq 4\nu(k+d)\log(k+d)$. Our proof is complete. ■

Lemma 26. *Let* $X \in \mathbb{R}^{k\times d}$ *be a probability matrix satisfying Assumption 3, and let* $Y_1, \ldots, Y_n$ *be i.i.d. one-hot samples with* $\mathbf{E}[Y_i] = X$*. Define the empirical estimation* $Y = \frac{1}{n}\sum_i Y_i$*, then*

$$\|Y - X\|_{\sigma,\infty} \leq \sqrt{\frac{16\nu \log((k+d)/\delta)}{n \min(k,d)}},$$

with probability exceeding $1 - \delta$ *when* $n \geq 4\nu(k+d)\log(k+d)$*.*

Proof of Lemma 26. Define $Z_i = (Y_i - X)/n$ for $1 \le i \le n$. We first note that $\|Z_i\|_{\sigma,\infty} \le (\|Y_i\|_{\sigma,\infty} + \|X\|_{\sigma,\infty})/n \le 2/n =: R$. Noting that $\mathbf{E}\, Y_i = X$ for all $1 \le i \le n$,

$$\mathbf{E}[Z_i Z_i^\top] = \mathbf{E}[Y_i Y_i^\top] - XX^\top.$$

Suppose $Y_i = e_j e_k^\top$, then $Y_i Y_i^\top = e_j e_j^\top$, similarly $Y_i^\top Y_i = e_k e_k^\top$. We then find

$$\mathbf{E}[Y_i Y_i^\top] = \begin{bmatrix} \sum_j X_{1,j} & 0 & \cdots & 0 \\ 0 & \sum_j X_{2,j} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sum_j X_{k,j} \end{bmatrix}$$

From Assumption 3, $\|XX^\top\|_{\sigma,\infty} \le \nu^2/\min^2(k,d)$ and $\|\mathbf{E}[Y_i Y_i^\top]\|_{\sigma,\infty} \le \nu/\min(k,d)$. We are now able to upper bound the variance statistic,

$$\sigma^2 := \max\left\{ \left\|\sum\nolimits_i \mathbf{E}[Z_i Z_i^\top]\right\|_{\sigma,\infty}, \left\|\sum\nolimits_i \mathbf{E}[Z_i^\top Z_i]\right\|_{\sigma,\infty} \right\} \le \frac{2\nu}{n\min(k,d)}.$$

Then from Lemma 30,

$$\mathbf{Pr}\left\{ \|Y - X\|_{\sigma,\infty} \ge t \right\} \le (k+d) \cdot \exp\left(\frac{-t^2/2}{\sigma^2 + Rt/3} \right).$$

After some algebraic manipulations,

$$\|Y - X\|_{\sigma,\infty} \le \sqrt{\frac{16\nu \log((k+d)/\delta)}{n\min(k,d)}},$$

with probability exceeding $1-\delta$ when $n \ge 4\nu(k+d)\log(k+d)$. ■

B Mathematical tools

Lemma 27 ([Gil52, Var57])**.** *The maximal size of a q-ary code of block length n and Hamming distance pn satisfies,*

$$A_q(n, pn) \ge q^{n(1-h_q(p))},$$

when $q \ge 2$ and $p \in [0, 1-1/q]$ where $h_q(x) = x\log_q(q-1) - x\log_q(x) - (1-x)\log_q(1-x)$.

Lemma 28 ([M+89])**.** *Suppose there are constants $c_1, \ldots, c_n$, such that for all $x_1 \in \mathcal{X}_1, \ldots, x_n \in \mathcal{X}_n$,*

$$\sup_{x_i' \in \mathcal{X}_i} |f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \le c_i.$$

Then for independent random variables $X_1 \in \mathcal{X}_1, \ldots, X_n \in \mathcal{X}_n$,

$$\mathbf{Pr}\left(|f(X_1, \ldots, X_n) - \mathbf{E}\left[f(X_1, \ldots, X_n)\right]| \ge \varepsilon \right) \le 2\exp\left(-2\varepsilon^2 / \sum\nolimits_i c_i^2 \right).$$

Lemma 29. *Given matrices $X, Y \in \mathbb{R}^{k\times d}$,*

$$\|X\|_{\sigma,1} - \|Y\|_{\sigma,1} \le \|\mathcal{P}_X(X - Y)\|_{\sigma,1}.$$

Proof of Lemma 29. We work through the inequality directly.

$$\|X\|_{\sigma,1} - \|Y\|_{\sigma,1} \le \|X\|_{\sigma,1} - \|\mathcal{P}_X Y\|_{\sigma,1} \le \|X - \mathcal{P}_X Y\|_{\sigma,1} = \|\mathcal{P}_X(X-Y)\|_{\sigma,1},$$

where the first inequality follows from the following manipulation,

$$\|\mathcal{P}_X Y\|_{\sigma,1} = \sup_{\|Z\|_{\sigma,\infty}\le 1} \langle \mathcal{P}_X Y, Z\rangle = \sup_{\|Z\|_{\sigma,\infty}\le 1} \langle Y, \mathcal{P}_X Z\rangle \le \|Y\|_{\sigma,1} \sup_{\|Z\|_{\sigma,\infty}\le 1} \|\mathcal{P}_X Z\|_{\sigma,\infty} \le \|Y\|_{\sigma,1},$$

and the second inequality follows from the reverse triangle inequality. ■

Lemma 30 ([Tro12])**.** *Let $Z_1, \ldots, Z_n$ be i.i.d random matrices in $\mathbb{R}^{k \times d}$. Suppose $\mathbf{E}\, Z_i = 0$ and $\|Z_i\|_{\sigma,\infty} \leq R$ almost surely for $1 \leq i \leq n$. Define the variance parameter as*

$$\sigma^2 \coloneqq \max\left\{ \left\| \sum_i \mathbf{E}[Z_i Z_i^\top] \right\|_{\sigma,\infty}, \left\| \sum_i \mathbf{E}[Z_i^\top Z_i] \right\|_{\sigma,\infty} \right\}.$$

Then the matrix Bernstein inequality states

$$\mathbf{Pr}\left\{ \left\| \sum_i Z_i \right\|_{\sigma,\infty} \geq t \right\} \leq (k+d) \cdot \exp\left(\frac{-t^2/2}{\sigma^2 + Rt/3} \right).$$