Low-rank matrix completion and denoising under Poisson noise

Andrew D. McRae   Mark A. Davenport
Georgia Institute of Technology

July 2, 2019

Abstract

This paper considers the problem of estimating a low-rank matrix from the observation of all, or a subset, of its entries in the presence of Poisson noise. When we observe all the entries, this is a problem of *matrix denoising*; when we observe only a subset of the entries, this is a problem of *matrix completion.* In both cases, we exploit an assumption that the underlying matrix is *low-rank.* Specifically, we analyze several estimators, including a constrained nuclear-norm minimization program, nuclear-norm regularized least squares, and a nonconvex constrained low-rank optimization problem. We show that for all three estimators, with high probability, we have an upper error bound (in the Frobenius norm error metric) that depends on the matrix rank, the fraction of the elements observed, and maximal row and column sums of the true matrix. We furthermore show that the above results are minimax optimal (within a universal constant) in classes of matrices with low rank and bounded row and column sums. We also extend these results to handle the case of matrix multinomial denoising and completion.

1   Introduction

1.1   Low-rank models for count data

In this paper, we consider the problem of estimating a non-negative matrix $M \in \mathbf{R}^{m \times n}$ given independent observations distributed according to Poisson(M_{ij}) for $(i, j) \in \Omega$, where Ω is a (not necessarily strict) subset of $\{1, \ldots, m\} \times \{1, \ldots, n\}$. If we do not make an observation for every entry of the matrix, the recovery problem is, in general, ill-posed in the absence of any additional assumptions on the underlying matrix. A common assumption for this type of problem is that the unknown matrix M is *low-rank*; i.e., the dimension of the spans of the columns and rows of M is much smaller than the actual numbers of columns and rows. This assumption greatly reduces the number of degrees of freedom in the model, making the recovery problem far more tractable. Note that even if we do observe every entry, we can still exploit the structure of the model to reduce the error due to noise.

While the problems of matrix completion and denoising have received a significant amount of attention in the settings of Gaussian noise and of small, bounded (in ℓ_2) perturbations (e.g., [1]–[3]), Poisson noise models have received comparatively less attention. In this paper focus primarily on the Poisson model, but we also examine closely-related multinomial models; this includes the case in which we have a single probability distribution over matrix coordinates (for which our result is a corollary of our Poisson results) as well as the case in which we make independent multinomial observations of matrix rows. Collectively, these models re often natural in applications where the observations arise via some form of counting process. The ability to recover (or de-noise) a low-rank signal from noisy, count-based observations is useful in many situations. We briefly mention two examples.

One potential application area involves imaging systems. This includes conventional cameras (which often suffer from noise in low light or with short exposures), but also 3-D imaging methods such as X-ray computed tomography (CT) and positron emission tomography (PET), which, in medical imaging, would greatly benefit from an improved noise/radiation dose tradeoff. In these scenarios, the Poisson noise model is natural because the observations consists of counts of particle (e.g., photon) arrivals at a detector. In many of these settings, such as when observing a periodic or slowly-varying sequence of images, a low-rank assumption on the underlying data is natural (see, e.g., [4] for an overview of low-rank modeling in image applications).

Another important application is topic modeling, which is a common form of dimensionality reduction for text documents. In this case, our observations consist of counts of word occurrences in a corpus of documents. If we suppose that these documents can be decomposed according to a small set of topics, and that within each topic documents will exhibit similar word occurrence counts, then a low-rank assumption on the word-frequency matrix is natural. For example, the popular PLSI model [5] uses a multinomial probability model parameterized by a low-rank matrix, but we note that existing algorithms have no performance guarantees.[1]

Low-rank models are also popular in nonnegative matrix factorization, which is commonly applied in a range of contexts where count data is common. A wide variety of such models have been developed, along with inference algorithms such as expectation maximization, variational Bayes, and Markov chain Monte Carlo [6]–[9]. These models and algorithms have been applied to many tasks, especially recommendation systems [10], [11]. However, the algorithms used are nonconvex, and there is little in the way of theoretical guarantees for their performance in the Poisson or multinomial setting.

1.2 Summary of main results for Poisson noise

In our analysis, we assume a Bernoulli sampling of the matrix entries: i.e., the events $\{(i,j) \in \Omega\}$ are independent with probability $p \in (0,1]$, and the observed Poisson random variables are independent conditioned on Ω. Note that taking $p = 1$ handles the case in which we observe every entry of the matrix. Below, for a matrix A, $\|A\|_*$, $\|A\|$, and $\|A\|_F$ denote the nuclear norm, operator norm, and Frobenius norm of A, respectively.

Let $\mathcal{A}_\Omega \colon \mathbf{R}^{m \times n} \to \mathbf{R}^\Omega$ denote the entry-wise sampling operator given by $(\mathcal{A}_\Omega(Z))(i,j) = Z_{ij}$ for $(i,j) \in \Omega$. Given observations $X \sim \mathrm{Poisson}(\mathcal{A}_\Omega(M))$, we consider several different estimators with similar theoretical properties. The first can be interpreted as a matrix version of the *Dantzig selector* [12]:

$$\widehat{M}^{(1)} = \operatorname*{argmin}_{M' \in [0,\infty)^{m \times n}} \|M'\|_* \text{ s.t. } \|\mathcal{A}_\Omega^*(X) - pM'\| \le \delta, \tag{1}$$

where $\delta > 0$ is a parameter which we will see how to set later. The second is a nuclear-norm-regularized least-squares type estimator:

$$\widehat{M}^{(2)} = \operatorname*{argmin}_{M' \in [0,\infty)^{m \times n}} \|\mathcal{A}_\Omega^*(X) - pM'\|_F^2 + \lambda \|M'\|_*, \tag{2}$$

where $\lambda > 0$ is another parameter which we will set. The third is least-squares under an exact low-rank constraint:

$$\widehat{M}^{(3)} = \operatorname*{argmin}_{M' \in [0,\infty)^{m \times n}} \|\mathcal{A}_\Omega^*(X) - pM'\|_F \text{ s.t. } \mathrm{rank}(M') \le r, \tag{3}$$

---

[1]We note that although PLSI assumes random word counts per document, it might be much more natural in some situations to make the number of words in each document fixed, which would require the independent-rows model that we also examine.

where r is an upper bound on the rank of the true rate matrix M. This problem is not convex (and is, in general, hard to solve directly while respecting nonnegativity constraints), but we will see later how we can address this issue without affecting its theoretical properties.

Theorem 1, which is the main result of Section 2.1, states that, if M has rank r, and hyperparameters are properly chosen, each of the estimators $\{\widehat{M}^{(i)}\}_{i=1}^3$ satisfies, with high probability,

$$\|M - \widehat{M}^{(i)}\|_F \lesssim \sqrt{\frac{r}{p}}\widetilde{\sigma}(M) + \text{logarithmic terms,} \tag{4}$$

where

$$\widetilde{\sigma}(M) = \max_i \sqrt{\sum_j M_{ij} + (1-p)M_{ij}^2} + \max_j \sqrt{\sum_i M_{ij} + (1-p)M_{ij}^2}.$$

In many situations (see Section 3.3), the logarithmic terms are negligible, so we can approximate this result by the bound

$$\|M - \widehat{M}^{(i)}\|_F \lesssim \sqrt{\frac{r}{p}}\widetilde{\sigma}(M). \tag{5}$$

Section 3 uses two standard methods to find lower bounds on the minimax risk of any estimator in classes of matrices with bounded row and column sums. These results (Theorems 3 and 4) can be summarized as follows: over all nonnegative matrices M such that $\text{rank}(M) \leq r$, and $\widetilde{\sigma}(M) \leq \sigma$, we have

$$\inf_{\widehat{M}} \sup_M \mathbf{E}_M \|M - \widehat{M}\|_F \gtrsim \sqrt{\frac{r}{p}}\sigma.$$

Thus, Theorem 1 is optimal (up to a multiplicative constant and an additive logarithmic factor) for this class of matrices.

To gain a more intuitive understanding of our result, it is helpful to examine the formula for $\widetilde{\sigma}$. For simplicity, assume, without loss of generality, that the row sums dominate the column sums, so that

$$\widetilde{\sigma} \approx \max_i \sqrt{\sum_j M_{ij} + (1-p)M_{ij}^2}.$$

The two terms inside the sum have different roles. The first term (M_{ij}) corresponds to the variance of the Poisson random variables. Indeed, if we take $p = 1$, this is the only term, so our result has the form

$$\|M - \widehat{M}^{(i)}\|_F \lesssim \sqrt{r}\left(\max_i \sqrt{\sum_j M_{ij}}\right).$$

If we do not impose any structure on the model, the maximum likelihood (and least-squares) estimate is $\widehat{M}^{\text{MLE}} = X$, which has risk

$$\mathbf{E}\|M - \widehat{M}^{\text{MLE}}\|_F = \sqrt{\sum_{i,j} M_{ij}}.$$

If every row of M has approximately the same sum, estimators defined above improve on $\widehat{M}^{\text{MLE}}$ (in squared Frobenius error) by a factor of approximately r/n. If the sums of the rows of M differ significantly, the improvement is smaller. However, this should not be too surprising — if the variance in the problem is already concentrated into a smaller sub-matrix, we are effectively solving a smaller problem, and hence the low-rank assumption is less restrictive and, therefore, less beneficial.

The second term (of the form $(1-p)M_{ij}^2$) in the formula for $\widetilde{\sigma}$ corresponds to the inherent difficulty in estimating the values of a matrix due to the fact that we do not observe every entry. This term in the lower bound applies regardless of the noise model, even when there is no noise. This might seem to contradict existing exact noiseless matrix completion results, but we note here that such results make stronger assumptions (incoherence of the row and column spaces) beyond what we are assuming here. In fact, the matrices used in the proof of Theorem 4 are highly coherent.

Although this second error term is necessary for general matrices, an interesting open problem is whether it could be entirely removed (leaving only the variance term) when we assume additional structure (such as incoherence) on the true rate matrix. Such a result would be a bridge between existing noisy and noiseless matrix completion literature; the existence of exact completion for the noiseless case implies that current results for the noisy case (including this paper) become highly suboptimal when the signal-to-noise ratio goes to infinity. An exception is [3], but we note that this approach is not without its own drawbacks as this approach leads to error rates which are suboptimal with respect to the rank r.

1.3   Summary of main results for multinomial denoising

We can also derive an interesting result on multinomial matrix denoising as a corollary of our result on Poisson denoising. If P is a non-negative $m \times n$ matrix such that $\sum_{i,j} P_{ij} = 1$, and we independently sample N objects according to the probabilities contained in P, the number of times each entry of P is sampled (which we can denote by an $m \times n$ count matrix X) has a (matrix) multinomial distribution. Our results on Poisson denoising apply by considering a multinomial distribution to be a vector of independent Poisson variables conditioned on its sum. Corollary 1 in Section 2.2 shows that if P has rank r, and none of the row and column sums of P are too large, there is an estimator $\widehat{P}$ of P (which could be defined similarly to any of the estimators above) such that

$$\|\widehat{P} - P\|_F \lesssim \sqrt{\frac{r}{N(m \wedge n)}}$$

One can easily check that, for the maximum likelihood estimator $\widehat{P}^{\mathrm{MLE}} = X/N$, $\mathbf{E}\|\widehat{P}^{\mathrm{MLE}} - P\|_F^2 \approx N^{-1}$, so our result (approximately) reduces the squared error by a factor of $r/(m \wedge n)$, which is the effective rank deficiency.

For a more complete exploration of multinomial matrix denoising, we also consider a model in which our observations are independent multinomial samples from rows of a low-rank matrix. Concretely, our observations are now a matrix X whose *rows*, which we denote $\{X_i\}_{i=1}^m$, are independent and distributed according to $X_i \sim \mathrm{Multinomial}(p_i, N_i)$, where $\{p_i\}_{i=1}^m$ are the rows of a rank-r $m \times n$ matrix P. Theorem 2 states that, under mild conditions on the sums of *columns* of P, there is an estimator $\widehat{P}$ (defined similarly to those above) such that, with high probability,

$$\|D^{1/2}(\widehat{P} - P)\|_F \lesssim \sqrt{r \log(m + n)},$$

where $D = \mathrm{diag}(N_1, \ldots, N_m)$. It is easily checked that the maximum likelihood estimator $\widehat{P}_{\mathrm{MLE}} = D^{-1}X$ has expected error $\mathbf{E}\|D^{1/2}(\widehat{P}^{\mathrm{MLE}} - P)\|_F^2 \approx m$, so we again get a reduction in (squared) error that is (approximately, modulo a logarithmic factor) proportional to the reduction in degrees of freedom.

We do not analyze the multinomial estimation problems from a minimax risk standpoint, but, due to the similarities between the Poisson and multinomial distributions, we suspect that one could find similar matching lower bounds in a similar manner to the Poisson case.

1.4 Computation and implications for general noisy matrix completion

Note that the two explicitly least-squares estimators $\widehat{M}^{(2)}$ and $\widehat{M}^{(3)}$ would be very easy to compute if we discarded the nonnegativity constraint; the we would take a singular value decomposition of the matrix $\mathcal{A}^*_\Omega(X)$, and then either do singular value soft thresholding (for $\widehat{M}^{(2)}$) or singular value truncation (for $\widehat{M}^{(3)}$).

We claim that ignoring the nonnegativity constraint does not change our analysis or the resulting error bounds at all; this constraint does not appear anywhere in the proof of Theorem 1. Therefore, if computational ability is a limiting factor, we could simply take the more efficient approach of solving without any nonnegativity constraints, and the error bounds we have presented will still apply. Projecting the result onto any convex constraint set that contains M can then only improve the performance.

We also note here that although we have chosen to focus on Poisson noise, our approach is fairly general and could apply to other types of noise. Indeed, if M were an arbitrary (not necessarily nonnegative) matrix, and we make observations of the form $M_{ij} + \xi_{ij}$ for $(i,j) \in \Omega$, and the ξ_{ij}'s are zero-mean noise variables with reasonably light tails, we could adapt our arguments to show that, for each of our three estimators,

$$\|M - \widehat{M}\|_F \lesssim \sqrt{\frac{r}{p}} \left(\max_i \sqrt{\sum_j \mathrm{var}(\xi_{ij}) + (1-p)M_{ij}^2} + \max_j \sqrt{\sum_i \mathrm{var}(\xi_{ij}) + (1-p)M_{ij}^2} \right).$$

The lower bound Theorem 4 is completely independent of the noise distribution; a version of Theorem 3 could be proved for many common distributions.

The above two points have some interesting implications for general matrix completion with noise. Many of the existing algorithms, such as low-rank factorization [13], iterative imputation [14], or the many other algorithms, including (1), that can be expressed as semidefinite programs, are fairly complex. Our results suggest that, not only do simple SVD-based algorithms have theoretical properties that are just as good as current state-of-the-art guarantees for more complex algorithms, but that, in a minimax error sense, *it is impossible to do any better*.

This realization does not, however, imply that there is no value to more sophisticated algorithms. As mentioned earlier, how well we can exploit incoherence in *noisy* matrix completion remains an important open question, and the matrices used in the proof of the minimax lower bound Theorem 4 are highly coherent. Therefore, it is likely that more sophisticated algorithms are still beneficial when trying to recover non-pathological (i.e., incoherent) matrices.

1.5 Comparison to prior work

There are several categories of existing literature to which we can compare our results. Some papers explicitly consider Poisson noise, using a maximum-likelihood framework. Cao and Xie [15] consider nuclear-norm penalized maximum likelihood for matrices contained in a nuclear norm ball (rather than exactly low-rank matrices). This approach uses an empirical process argument to bound the Kullback-Leibler divergence between the true and predicted distributions. This argument requires a Lipschitz condition on the log-likelihood function, which, for the Poisson distribution, requires imposing a lower bound on the rates. Soni, Jain, Haupt, *et al.* [16] and Soni and Haupt [17] consider a penalized maximum likelihood estimator from a carefully-chosen finite set of candidates (which is exponentially large in the size of the problem and hence computationally intractable). The matrices considered have a non-negative low-rank factorization (with a particular emphasis on the case when one factor is sparse). They use an information-theoretic argument to bound the expected error in

terms of Bhattacharyya distance. The result of [16], which applies to matrix completion, requires imposing a lower bound on the rates, while that of [17], which considers only denoising, does not. All three papers find an upper bound on Frobenius error in terms of the statistical error metrics that they originally bound.

Other, more general approaches, are designed specifically with Frobenius-norm error in mind. One class of methods uses "restricted strong convexity" arguments, which were introduced by Negahban and Wainwright [1]. These methods rely on approximating the Frobenius norm in certain restricted classes of matrices (in which the error matrix must fall) using only samples of the entries. These methods lead to simple and elegant proofs, but the concentration inequalities on which they rely require imposing uniform upper bounds in magnitude on both the true matrix entries and the estimator entries. Other recent papers which use this type of argument include [18]–[21].

Another class of methods is direct singular value thresholding. Papers with this approach include [2] (which inspired our approach) as well as [22]–[24]. These methods are very simple and lend themselves to simple proofs.

An interesting blend of techniques can be seen in the works of Lafond [25] and Gunasekar, Ravikumar, and Ghosh [26], who combine some of the general approaches mentioned above with maximum likelihood estimation for exponential families of distributions. These methods, like those in [15] and [16], are difficult to apply to the Poisson distribution without imposing a lower bound on rates because, as the mean λ of the distribution goes to 0, the "natural parameter" $\log \lambda$ goes to $-\infty$, whereas the general methods used require parameters to be bounded. They also require (approximate) low rank in the matrix of natural parameters. In the Poisson case, this is equivalent to assuming a bound on the rank of the matrix $[\log M_{ij}]$ of elementwise logarithms of the means, which is somewhat non-standard, and certainly not the same as bounding the rank of the original matrix M.

Most of the papers mentioned above do not find error bounds which explicitly depend on the "true" rate matrix; rather, they find uniform upper bounds for classes of structured matrices with uniform upper (and, sometimes, lower) bounds on the entries. To compare our results directly to this literature, we consider what we obtain when we only impose a uniform upper and lower bounds (by, say $\lambda_{\max}$ and $\lambda_{\min}$) on the matrix entries. The approximate bound of (5) reduces to

$$\|\widehat{M} - M\|_F^2 \lesssim (\lambda_{\max} + (1-p)\lambda_{\max}^2)\frac{rm}{p},$$

where we have assumed, without loss of generality, that $m \geq n$. Previous results show similar error rates in terms of matrix dimensions for exactly low-rank matrices. For example, [16] establishes a bound of

$$\mathbf{E}\|\widehat{M} - M\|_F^2 \lesssim \frac{\lambda_{\max}^3}{\lambda_{\min}}\frac{rm}{p}\log m,$$

which provides a similar dependence on r, m, and p, but with an additional logarithmic term and a worse dependence on the minimum and maximum matrix values. In a slightly different setting, [15] shows that for matrices in the nuclear norm ball of radius $\lambda_{\max}\sqrt{rmn}$ (which is a convex relaxation of the exact low-rank constraint), we instead obtain (ignoring logarithmic terms and a complicated but severe dependence on $\lambda_{\max}$ and $\lambda_{\min}$) an error bound of

$$\|\widehat{M} - M\|_F^2 \lesssim \frac{\sqrt{rnm}}{\sqrt{p}},$$

where p is now the number of samples for entry in a uniform-at-random sampling model. The different dependence on r and p is interesting, but, if one compares it to results in linear regression

over ℓ_1 balls (see, e.g., [27]), the rate given is perhaps not surprising. To compare to some of the more general methods mentioned above, we note that, if we consider the generalization of our method mentioned in Section 1.4, and we assume a uniform upper bound on the magnitudes of matrix elements and the noise variances, our results are comparable to [18] (albeit under less-strict assumptions).

As noted earlier, our paper uses fairly general matrix completion methods. For the Frobenius norm error metric, this gives us an advantage over more distribution-specific approaches such as [15], [16], [25], [26], in part because we do not have to approximate the Frobenius-norm error by a statistical divergence measure or by a norm in a transformed parameter space. Our results also do not suffer from the fact that a Poisson distribution's likelihood function is ill-conditioned for very small rates. In addition, our results avoid a multiplicative logarithmic factor that appears in much of the previous literature (replacing it with an additive factor that is often negligible); this achievement (which also appears in [18]) is almost entirely due to the use of recent results in bounding the operator norm of a random matrix (such as [28]).

Finally, much of the previous literature in the Poisson case (from those mentioned above, [15], [16], [25]) finds lower bounds on minimax risk in certain classes of matrices. Although these lower bounds have the same large-scale error rate (in terms of the rank and dimensions of the matrix and the number of samples) as the corresponding upper bounds, they differ from the upper bounds by factors that are logarithmic in the problem size and that depend on the ratio of largest to smallest allowable rates. To our knowledge, the results in this paper are the first for noisy low-rank matrix completion in which the minimax rate for large classes of matrices is found to within a universal constant.

We are aware of much less theoretical work for low-rank multinomial denoising. Most of the existing work in the topic modeling literature does not attempt to find theoretical guarantees of performance. One recent work that is worth noting is [29]. This paper shows that, in the case of a matrix multinomial distribution, one can achieve a tight error bound in ℓ_1 distance (sum of absolute values of entries) for certain factorizable probability matrices using $O(mr^4)$ samples. It is difficult to compare this directly to our result, since it is in the much stronger ℓ_1 norm, but we note that, in a similar fashion as many of the results on Poisson observations, this paper also relies on lower bounding certain sums of entries in the factor matrices.

We add a final caveat to our results by noting that $\|\widehat{M} - M\|_F$ might not always be the most appropriate error metric; for example, there is a much larger difference qualitatively (and quantitatively, if we use an appropriate statistical divergence) between Poisson distributions of means 0 and 10 than between Poisson distributions of means 100 and 110. We see a similar disconnect between squared error and other probabilistic metrics in the case of the multinomial distribution. Further investigation of distribution-specific methods (such as maximum likelihood) that yield bounds in more statistically-motivated metrics is thus certainly warranted.

1.6 Paper outline

The remainder of this paper is organized as follows. Section 2 contains the the formal statements and proofs of the upper bounds on error in this paper. Section 3 contains the statements and proofs of two separate minimax lower bounds which, when combined, yield the matching lower bound to (5). Section 3.3 also discusses briefly some situations where the approximation of (5) is accurate, and thus the upper and lower bounds match within a universal constant.

2 Upper bounds

2.1 Poisson noise

This section is dedicated to proving the main result of this paper, which is the following theorem.

Theorem 1. *Let M be a non-negative $m \times n$ matrix with rank r. Let $\lambda_{max} = \max_{ij} M_{ij}$, and let*

$$\widetilde{\sigma}(M) = \max_i \sqrt{\sum_j M_{ij} + (1-p)M_{ij}^2} + \max_j \sqrt{\sum_i M_{ij} + (1-p)M_{ij}^2}.$$

Suppose $\Omega \subset \{1, \dots, m\} \times \{1, \dots, n\}$ is chosen according to a Bernoulli sampling model with sampling probability p, and suppose, conditionally on Ω, $X \sim \text{Poisson}(\mathcal{A}_\Omega(M))$. Set $\epsilon \in (0, 1/2)$, and let

$$A(M, p, \epsilon) = 2\sqrt{p}\widetilde{\sigma}(M) + \frac{8\epsilon}{\sqrt{mn}} + C \max\left\{\lambda_{\max}, 4\log\frac{2mn}{\epsilon}\right\}\sqrt{\log\frac{m \vee n}{\epsilon}}, \tag{6}$$

where C is a universal constant.

Then, with probability at least $1 - 2\epsilon$, if $\delta \geq A(M, p, \epsilon)$ and $\lambda \geq 2pA(M, p, \epsilon)$, we also have

$$\|\widehat{M}^{(1)} - M\|_F \leq \frac{4\sqrt{2r}\delta}{p}$$

and

$$\|\widehat{M}^{(2)} - M\|_F \leq \frac{2\sqrt{2r}\lambda}{p^2}.$$

Moreover, we also have that with probability at least $1 - 2\epsilon$,

$$\|\widehat{M}^{(3)} - M\|_F \leq \frac{2\sqrt{2r}}{p}A(M, p, \epsilon).$$

The result follows from a series of lemmas. The first steps in upper bounding the error are the following (deterministic) results.

Lemma 1. *Suppose M is a rank-r matrix such that $\|\mathcal{A}_\Omega^*(X) - pM\| \leq \delta = \frac{\lambda}{2p}$. Then, for $i \in \{1, 2\}$,*

$$\|\widehat{M}^{(i)} - M\|_F \leq \frac{4\sqrt{2r}\delta}{p} = \frac{2\sqrt{2r}\lambda}{p^2}. \tag{7}$$

Lemma 2. *Suppose M is a rank-r matrix. Then*

$$\|\widehat{M}^{(3)} - M\|_F \leq \frac{2\sqrt{2r}}{p}\|\mathcal{A}_\Omega^*(X) - pM\|.$$

Proof of Lemma 1. Let $M = U\Sigma V^*$ be the singular value decomposition of M, where $U \in \mathbf{R}^{m \times r}$ and $V \in \mathbf{R}^{n \times r}$ are such that $U^*U = V^*V = I_r$, and Σ is an $r \times r$ diagonal matrix with positive entries on the diagonal. Let T be the subspace of $\mathbf{R}^{m \times n}$ spanned by matrices of the form UA and BV^T for arbitrary matrices $A \in \mathbf{R}^{r \times n}$ and $B \in \mathbf{R}^{m \times r}$. We denote by $\mathcal{P}_T$ and $\mathcal{P}_{T^\perp}$, respectively, the orthogonal projections onto T and its orthogonal complement $T^\perp$.

Denote $H^{(1)} = \widehat{M}^{(1)} - M$. Because M is feasible, and the nuclear norm is a convex function, we have

$$0 \geq \|\widehat{M}^\delta\|_* - \|M\|_* \geq \langle H^{(1)}, Z\rangle,$$

where $Z \in \partial \|M\|_*$ is any subgradient of the nuclear norm function at the point M. Such a subgradient must have the form

$$Z = UV^* + \mathcal{P}_{T^\perp}(W),$$

where W is an arbitrary matrix with $\|W\| \le 1$. By the duality of the nuclear norm and operator norm, we can choose W so that $\langle W, \mathcal{P}_{T^\perp}(H^{(1)}) \rangle = \|\mathcal{P}_{T^\perp}(H^{(1)})\|_*$. We then have

$$
\begin{aligned}
0 &\ge \langle H^{(1)}, UV^* + \mathcal{P}_{T^\perp}(W) \rangle \\
&= \langle \mathcal{P}_T(H^{(1)}), UV^* \rangle + \|\mathcal{P}_{T^\perp}(H^{(1)})\|_* \\
&\ge -\|\mathcal{P}_T(H^{(1)})\|_* + \|\mathcal{P}_{T^\perp}(H^{(1)})\|_*,
\end{aligned}
$$

where the last inequality follows from the fact that $\|UV^*\| = 1$. We therefore have

$$\|\mathcal{P}_{T^\perp}(H^{(1)})\|_* \le \|\mathcal{P}_T(H^{(1)})\|_*.$$

Hence

$$
\begin{aligned}
\|H^{(1)}\|_* &\le 2\|\mathcal{P}_T(H^{(1)})\|_* \\
&\le 2\sqrt{2r}\|\mathcal{P}_T(H^{(1)})\|_F \\
&\le 2\sqrt{2r}\|(H^{(1)})\|_F,
\end{aligned}
$$

where the second inequality follows from the fact that any element of T has rank at most $2r$.

By the triangle inequality and the homogeneity of the operator norm, we also have

$$\|H^{(1)}\| \le \frac{2\delta}{p}.$$

Thus

$$
\begin{aligned}
\|H^{(1)}\|_F^2 &= \langle H^{(1)}, H^{(1)} \rangle \\
&\le \|H^{(1)}\|\|H^{(1)}\|_* \\
&\le \frac{4\sqrt{2r}\delta}{p}\|H^{(1)}\|_F,
\end{aligned}
$$

and the first part of the result immediately follows.

The proof of the second part is similar; letting $H^{(2)} = \widehat{M}^{(2)} - M$, we now have, by the optimality of $\widehat{M}$,

$$
\begin{aligned}
0 &\ge \|\mathcal{A}_\Omega^*(X) - p\widehat{M}^{(2)}\|_F^2 + \lambda\|\widehat{M}^{(2)}\|_* - \left(\|\mathcal{A}_\Omega^*(X) - pM\|_F^2 + \lambda\|M\|_* \right) \\
&= p^2\|\widehat{M}^{(2)}\|_F^2 - p^2\|M\|_F^2 + p\langle \mathcal{A}_\Omega^*(X), M - \widehat{M}^{(2)} \rangle + \lambda\left(\|\widehat{M}^{(2)}\|_* - \|M\|_* \right) \\
&= p^2\|H^{(2)}\|_F^2 + 2p\langle \mathcal{A}_\Omega^*(X) - pM, H^{(2)} \rangle + \lambda\left(\|\widehat{M}^{(2)}\|_* - \|M\|_* \right).
\end{aligned}
$$

Noting, as before, that $\|\widehat{M}^{(2)}\|_* - \|M\|_* \ge \|\mathcal{P}_{T^\perp}(H^{(2)})\|_* - \|\mathcal{P}_T(H^{(2)})\|_*$, and that

$$|\langle \mathcal{A}_\Omega^*(X) - pM, H^{(2)} \rangle| \le \|\mathcal{A}_\Omega^*(X) - pM\|\|H^{(2)}\|_* \le \frac{\lambda}{2p}\|H^{(2)}\|_*,$$

we have

$$\|H^{(2)}\|_F^2 \leq \frac{\lambda}{p^2}(\|H^{(2)}\|_* + \|\mathcal{P}_T(H^{(2)})\|_* - \|\mathcal{P}_{T^\perp}(H^{(2)})\|_*)$$

$$\leq \frac{2\lambda}{p^2}\|\mathcal{P}_T(H^{(2)})\|_*$$

$$\leq \frac{2\sqrt{2r}\lambda}{p^2}\|H^{(2)}\|_F.$$

$\square$

Proof of Lemma 2. Because both M and $\widehat{M}^{(3)}$ have rank at most r, $H^{(3)} = \widehat{M}^{(3)} - M$ has rank at most $2r$. The optimality of $\widehat{M}^{(3)}$ implies

$$0 \geq \|\mathcal{A}_\Omega^*(X) - p\widehat{M}^{(3)}\|_F^2 - \|\mathcal{A}_\Omega^*(X) - pM\|_F^2$$
$$= 2p\langle\mathcal{A}_\Omega^*(X) - pM, H^{(3)}\rangle + p^2\|H^{(3)}\|_F^2,$$

so

$$\|H^{(3)}\|_F^2 \leq \frac{2}{p}|\langle\mathcal{A}_\Omega^*(X) - pM, H^{(3)}\rangle|$$

$$\leq \frac{2}{p}\|\mathcal{A}_\Omega^*(X) - pM\|\|H^{(3)}\|_*$$

$$\leq \frac{2\sqrt{2r}}{p}\|\mathcal{A}_\Omega^*(X) - pM\|\|H^{(3)}\|_F.$$

$\square$

The remainder of the work is to show that $\|\mathcal{A}_\Omega^*(X) - pM\| \leq A(M, p, \epsilon)$ with probability at least $1 - 2\epsilon$. We will use the following fundamental lemma, which was originally proved by Bandeira and van Handel [28] and appears with a slightly improved constant in [30].

Lemma 3 (Theorem 4.9 and Remark 4.11 in [30])**.** *Let X be a random $m \times n$ matrix whose entries are independent, centered, and almost surely bounded in absolute value by a constant b. Let*

$$\sigma = \max_i \sqrt{\sum_j \mathbf{E}\, X_{ij}^2} + \max_j \sqrt{\sum_i \mathbf{E}\, X_{ij}^2}.$$

Then

$$\mathbf{P}(\|X\| \geq 2\sigma + t) \leq (m \vee n)\exp\left(-\frac{t^2}{C_0 b^2}\right),$$

where C_0 is a universal constant.

Poisson random variables are clearly unbounded, so Lemma 3 does not directly apply. The following technical lemma allows us to extend the result to the case of random variables with sub-exponential tails.

Lemma 4. *Let X be a random $m \times n$ matrix whose entries are independent and centered, and suppose that for some $v, t_0 > 0$, we have, for all $t \geq t_0$,*

$$\mathbf{P}(|X_{ij}| \geq t) \leq 2e^{-t/v}.$$

10

Let $\epsilon \in (0, 1/2)$, and let

$$K = \max \left\{ t_0, v \log \frac{2mn}{\epsilon} \right\}.$$

Then

$$\mathbf{P}\left(\|X\| \geq 2\sigma + \frac{\epsilon v}{\sqrt{mn}} + t \right) \leq (m \vee n) \exp \left(-\frac{t^2}{C_0 (2K)^2} \right) + \epsilon,$$

where σ and C_0 are the same as in Lemma 3.

Proof. First, note that, by a union bound,

$$\mathbf{P}\left(\max_{i,j} |X_{ij}| > K \right) \leq 2mne^{-K/v} \leq \epsilon.$$

Consider the truncation $X^K = [X_{ij}^K]$, where $X_{ij}^K = X_{ij} \mathbf{1}_{\{|X_{ij}| \leq K\}}$. Note that

$$
\begin{aligned}
|\mathbf{E}\, X_{ij}^K| &\leq \mathbf{E}|X_{ij}^K - X_{ij}| \\
&= \mathbf{E}|X_{ij}| \mathbf{1}_{\{|X_{ij}| > K\}} \\
&= \int_K^\infty \mathbf{P}(|X_{ij}| > t)\, dt \\
&\leq \int_K^\infty 2e^{-t/v}\, dt \\
&= 2ve^{-K/v} \\
&\leq \frac{\epsilon v}{mn} \\
&\leq K.
\end{aligned}
$$

Let $\widetilde{X}^K = X^K - \mathbf{E}\, X^K$ be the centered version of X^K. Clearly, $\mathbf{E}(\widetilde{X}_{ij}^K)^2 \leq \mathbf{E}\, X_{ij}^2$, and $|\widetilde{X}_{ij}^K| \leq 2K$. Then, by Lemma 3,

$$\mathbf{P}(\|\widetilde{X}^K\| \geq 2\sigma + t) \leq (m \vee n)e^{-t^2/C_0(2K)^2}.$$

Furthermore, with probability at least $1 - \epsilon$,

$$
\begin{aligned}
\|X\| = \|X^K\| &\leq \|\widetilde{X}^K\| + \|\mathbf{E}\, X^K\| \\
&\leq \|\widetilde{X}^K\| + \|\mathbf{E}\, X^K\|_F \\
&\leq \|\widetilde{X}^K\| + \frac{\epsilon v}{\sqrt{mn}},
\end{aligned}
$$

and the result follows. $\qquad\square$

To apply this result, we need a subexponential tail bound for the Poisson distribution.

Lemma 5. *Let $X \sim \text{Poisson}(\lambda)$. Then*

$$\mathbf{P}(X - \lambda \geq t) \leq \exp \left(-\frac{t^2}{2(\lambda + t/3)} \right).$$

For $t \geq \lambda$,

$$\mathbf{P}(X - \lambda \geq t) \leq e^{-3t/8}.$$

The first inequality can be established by approximating the Poisson distribution with mean λ as the sum of k Bernoulli random variables with mean λ/k, applying Bernstein's inequality, and taking $k \to \infty$. The idea for this argument was suggested by an exercise in [31].

Going back to our original problem, we need to bound the operator norm of $Z = \mathcal{A}_\Omega^*(X) - pM$. Note that since we are using a Bernoulli sampling model, the entries of Z are independent. Let $\lambda_{\max} = \max_{i,j} M_{ij}$. Note that for every (i,j), $\mathbf{E}\, Z_{ij} = 0$,

$$Z_{ij} \geq -pM_{ij} \geq -\lambda_{\max},$$

and, for $t \geq 2\lambda_{\max}$,

$$\mathbf{P}(Z_{ij} \geq t) \leq e^{-3(t-\lambda_{\max})/8} \leq e^{-3t/16} \leq e^{-t/8}.$$

Then, by Lemma 4, we have, for $\epsilon \in (0, 1/2)$,

$$\mathbf{P}\left(\|X\| \geq 2\sigma + \frac{8\epsilon}{\sqrt{mn}} + t \right) \leq (m \vee n) \exp\left(-\frac{t^2}{C_0 (2K)^2} \right) + \epsilon,$$

where

$$K = \max\left\{ 2\lambda_{\max}, 8\log\frac{2mn}{\epsilon} \right\},$$

and σ is defined as before. To calculate σ in terms of p and M, we note that

$$\mathrm{var}(Z_{ij}) = p\,\mathrm{var}(X_{ij}) + p(1-p)(\mathbf{E}\, X_{ij})^2$$
$$= pM_{ij} + p(1-p)M_{ij}^2.$$

Therefore, we can calculate

$$\sigma = \max_i \sqrt{\sum_j pM_{ij} + p(1-p)M_{ij}^2} + \max_j \sqrt{\sum_i pM_{ij} + p(1-p)M_{ij}^2} = \sqrt{p}\,\tilde{\sigma}.$$

2.2 Corollary on multinomial estimation

Here we prove a corollary of Theorem 1, showing how it implies a result for estimating the low-rank-matrix parameter of a matrix multinomial distribution. For the sake of brevity, we only consider the analogue of $\widehat{M}^{(1)}$.

Corollary 1. *Let P be a nonnegative $m \times n$ matrix with rank r such that $\sum_{i,j} P_{ij} = 1$. Suppose, furthermore, that we have*

$$\max_i \sum_j P_{ij} \leq \frac{a}{m}, \max_j \sum_i P_{ij} \leq \frac{b}{n}$$

for some constants $a, b \geq 1$, and that $\max_{i,j} P_{ij} \leq c$. Let N be a positive integer, and suppose that $X \sim \mathrm{Multinomial}(P, N)$. Let $\epsilon \in (0, 1)$, choose $\delta > 0$ such that

$$\delta \geq \frac{1}{N}\left(2\sqrt{N\left(\frac{a}{m} + \frac{b}{n}\right)} + \frac{4\epsilon}{e\sqrt{mnN}} + C\max\left\{ Nc, 4\log\frac{4emn\sqrt{N}}{\epsilon} \right\} \sqrt{\log\frac{2e\sqrt{N}(m \vee n)}{\epsilon}} \right),$$

and let

$$\widehat{P}^\delta(X) = \operatorname*{argmin}_{\substack{P' \in [0,1]^{m \times n} \\ \sum_{i,j} P_{ij} = 1}} \|P'\|_* \text{ s.t. } \|X - NP'\| \leq N\delta.$$

Then, with probability at least $1 - \epsilon$,

$$\|\widehat{P}^\delta - P\|_F \leq 4\sqrt{2r}\delta.$$

As in the Poisson case, there are many situations where the additive logarithmic term in the definition of δ is negligible, and we have

$$\|\widehat{P}^\delta - P\|_F \lesssim \sqrt{\frac{r}{N}\left(\frac{a}{m} + \frac{b}{n}\right)}.$$

Proof of Corollary 1. Suppose $Y \sim \text{Poisson}(NP)$. Let $\epsilon' = \epsilon/2e\sqrt{N}$. Define the event $A = \{\|\widehat{P}^\delta(Y) - P\|_F \leq 4\sqrt{2r}\delta\}$. Theorem 1 (with $p = 1$) implies that $\mathbf{P}(A) \geq 1 - 2\epsilon' = 1 - \epsilon/e\sqrt{N}$. Note that the additional constraints in the optimization problem do not affect this fact, since P, by definition, meets these constraints.

X has the same distribution as Y conditioned on $B = \{\sum_{i,j} Y_{ij} = N\}$, so it suffices to show that the probability of A conditioned on this event is at least $1 - \epsilon$. Indeed, note that $\sum_{i,j} Y_{ij} \sim \text{Poisson}(N)$, so

$$\mathbf{P}(B) = \frac{e^{-N}N^N}{N!} \geq \frac{1}{e\sqrt{N}}$$

by Stirling's approximation. Then, by Bayes' rule,

$$\begin{aligned}
\mathbf{P}(\|\widehat{P}^\delta(X) - P\|_F \geq 4\sqrt{2r}\delta) &= \mathbf{P}(A^c \mid B) \\
&= \frac{\mathbf{P}(A^c \cap B)}{\mathbf{P}(B)} \\
&\leq \frac{\mathbf{P}(A^c)}{\mathbf{P}(B)} \\
&\leq 2\epsilon' e\sqrt{N} \\
&= \epsilon
\end{aligned}$$

$\square$

2.3 Multinomial denoising with independent rows

We now consider a slightly different setting for multinomial estimation. Here, we consider a model in which we observe a collection of independent multinomial random variables, each of which has distribution parametrized by a row in a low-rank matrix.

Theorem 2. *Let $X_1, \ldots, X_m$ be independent multinomial random variables, with*

$$X_i \sim \text{Multinomial}(p_i, N_i),$$

where, for each i, $N_i \geq 1$ is an integer, and $p_i = (p_{i1}, \ldots, p_{in})$ is a vector of probabilities. Denote the matrices

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_m \end{bmatrix} \in \mathbf{R}^{m \times n}, \ P = \begin{bmatrix} p_1 \\ \vdots \\ p_{m,} \end{bmatrix} \in \mathbf{R}^{m \times n}, \ D = \begin{bmatrix} N_1 & & \\ & \ddots & \\ & & N_m \end{bmatrix} \in \mathbf{R}^{m \times m},$$

where each X_i and p_i are considered row vectors. Define the estimator

$$\widehat{P}^\delta = \underset{P' \in [0,1]^{m \times n}}{\text{argmin}} \ \|D^{1/2}P'\|_* \ \text{s.t.} \ \|D^{-1/2}(X - DP')\| \leq \delta, \ \sum_j P_{ij} = 1 \ \forall i \in \{1, \ldots, m\}.$$

13

Let $D_{\min} = \min_i D_i$, and choose δ such that

$$\delta \geq \max\left\{2\sqrt{\max\left\{1, \max_j \sum_i p_{ij}\right\}\log\frac{m+n}{\epsilon}}, \frac{4}{3\sqrt{D_{\min}}}\log\frac{m+n}{\epsilon}\right\}.$$

Then, with probability at least $1 - \epsilon$,

$$\|D^{1/2}(\widehat{P}^\delta - P)\|_F \leq 4\sqrt{2r}\delta,$$

where r is the rank of P.

Our approach uses the following matrix Bernstein inequality.

Lemma 6 ([32, Theorem 6.1.1]). *Suppose $Z = \sum_k S_k \in \mathbf{R}^{m \times n}$, where $\{S_k\}_k$ is a finite sequence of independent, zero-mean random matrices. Suppose that for some constant $L > 0$, for each k, $\|S_k\| \leq L$ almost surely. Let*

$$v = \max\{\|\mathbf{E}\, ZZ^T\|, \|\mathbf{E}\, Z^TZ\|\} = \max\left\{\left\|\sum_k \mathbf{E}\, S_k S_k^T\right\|, \left\|\sum_k \mathbf{E}\, S_k^T S_k\right\|\right\}.$$

Then, for $t \geq 0$,

$$\mathbf{P}(\|Z\| \geq t) \leq (m+n)\exp\left(\frac{-t^2}{2(v + Lt/3)}\right).$$

Proof of Theorem 2. On the event that $\|D^{-1/2}(X - DP)\| \leq \delta$, we have $\|D^{1/2}(\widehat{P}^\delta - P)\| \leq 2\delta$, and the result follows by the same steps as in the proof of Theorem 1.

To find a bound on the operator norm of $Z = D^{-1/2}(X - DP)$, we apply Lemma 6, noting that $D^{-1/2}(X - DP)$ is the sum of $\sum_{i=1}^m D_i$ random, zero-mean matrices of the form $\frac{1}{\sqrt{D_i}}e_i u_{ij}^T$, where e_i is the ith standard basis element in $\mathbf{R}^m$, and u_{ij} is a vector in $\mathbf{R}^n$ that is equal to e_k with probability p_{ik}.

One can verify that

$$\mathbf{E}\, ZZ^T = \text{diag}\left(\sum_{j=1}^n p_{1j}(1 - p_{1j}), \ldots, \sum_{j=1}^n p_{mj}(1 - p_{mj})\right) \preceq I_m,$$

where I_m denotes the $m \times m$ identity matrix, and that

$$\mathbf{E}\, Z^TZ = \sum_{i=1}^m \text{diag}(p_i) - p_i p_i^T \preceq \text{diag}\left(\sum_{i=1}^m p_{i1}(1 - p_{i1}), \ldots, \sum_{i=1}^m p_{in}(1 - p_{in})\right),$$

where $p_i = (p_{i1}, \ldots, p_{in})$ is the ith row of p. We can then take $v \leq \max\{1, \max_j \sum_i p_{ij}\}$. Clearly, each term in the sum has operator norm bounded above by $L = 1/\sqrt{D_{\min}}$.

Some algebraic manipulation of the result of Lemma 6 implies that, for $\epsilon \in (0, 1)$, we have, with probability at least $1 - \epsilon$,

$$\|Z\| \leq \max\left\{2\sqrt{v\log\frac{m+n}{\epsilon}}, \frac{4L}{3}\log\frac{m+n}{\epsilon}\right\}$$

$$\leq \max\left\{2\sqrt{\max\left\{1, \max_j \sum_i p_{ij}\right\}\log\frac{m+n}{\epsilon}}, \frac{4}{3\sqrt{D_{\min}}}\log\frac{m+n}{\epsilon}\right\},$$

which establishes the result.

$\square$

3 Minimax lower bounds

In this section, we show that the rate in (5) is optimal (within a multiplicative constant) in the sense of minimax risk. We do this in two parts; the two minimax lower bounds derived in the next two sections, when combined, match the rate in (5) is optimal.

3.1 First lower bound

In the Poisson upper error bound, $\widetilde{\sigma}$ is partially determined by the maximal row and column sums of the rate matrix M, which we can think of as the maximal variance of any row or column (without sampling a subset of the entries). Our first lower bound shows that we cannot improve on this term:

Theorem 3. *Let r, k, and ℓ, be positive integers, and take $m = rk$, $n = r\ell$. Let $\lambda_{\max} \geq 1/8\ell p$, set $\sigma_1^2 = k\lambda_{\max}$, and let*

$$
S = \left\{ M \in [0, \lambda_{\max}]^{m \times n} : \mathrm{rank}(M) \leq r, \ \sqrt{\max_i \sum_j M_{ij}} + \sqrt{\max_j \sum_i M_{ij}} \leq 2\sigma_1 \right\}.
$$

Then, under a Bernoulli sampling model with sampling probability p,

$$
\inf_{\widehat{M}} \sup_{M \in S_1} \mathbf{P}_M \left(\|\widehat{M} - M\|_F \geq \frac{\sqrt{r}\sigma_1}{8\sqrt{2p}} \right) \geq \frac{1}{2} - \frac{8\log 2}{m \vee n}.
$$

Proof. Assume, without loss of generality, that $k \geq \ell$. We use a variant of Fano's method. We first find a large hypercube of matrices, and then we use the fact that we can find a large subset that is well-separated.

For $i \in \{1, \ldots, m\}$, let

$$
A_i = \{\ell(q-1) + 1, \ldots \ell q\},
$$

where $q \in \{1, \ldots, r\}$ is the unique integer such that $i \in \{k(q-1) + 1, \ldots, kq\}$. For λ_0, λ_1 to be chosen later, we define, for $\theta \in \{0, 1\}^m$, the block-diagonal matrix $M_\theta \in S$ by

$$
(M_\theta)_{ij} = \begin{cases} \lambda_{\theta_i}, & j \in A_i, \\ 0, & \text{otherwise.} \end{cases}
$$

The nonzero elements of the ith row of M_θ are all either λ_0 or λ_1, depending on the value of θ_i.

By a combinatorial argument (see, e.g., [33]), one can show that there exists $\Theta \subset \{0, 1\}^m$ such that $\mathrm{card}(\Theta) \geq e^{m/8}$ and, for all distinct $\theta, \theta' \in \Theta$, the Hamming distance $d_H(\theta, \theta') \geq m/4$.

Take $\lambda_0 = \lambda_{\max}/2 - \delta$ and $\lambda_1 = \lambda_{\max}/2 + \delta$, where $\delta \leq \lambda_{\max}/2$ is a constant to be chosen later. Note that for all distinct $\theta, \theta' \in \Theta$, we have

$$
\|M_\theta - M_{\theta'}\|_F \geq \sqrt{m\ell}\delta.
$$

We denote by P_θ the distribution of $\mathcal{P}_\Omega(X)$ when $X \sim \mathrm{Poisson}(M_\theta)$, and Ω is independently chosen from the Bernoulli sampling model with parameter p. From Fano's inequality from information theory, one can derive (see, e.g., [33]) a lower bound on the probability of an estimator's error

exceeding half the distance between points indexed by Θ:

$$\inf_{\widehat{M}} \sup_{M \in S} \mathbf{P}_M \left(\|\widehat{M} - M\|_F \geq \frac{\sqrt{m\ell}\delta}{2} \right) \geq p_e$$

$$:= \inf_{\phi} \sup_{\theta \in \Theta} \mathbf{P}(\phi(\mathcal{P}_\Omega(X)) \neq \theta)$$

$$\geq 1 - \frac{\sup_{\theta \in \Theta} \mathrm{D_{KL}}(P_\theta \,\|\, Q) + \log 2}{\log \mathrm{card}(\Theta)},$$

where ϕ denotes a test taking values in Θ, and Q is any probability distribution on $\mathbf{R}^{m \times n}$. We take Q to be the distribution generated in the same manner as each P_θ, simply with λ_1 and λ_2 replaced by $\lambda_{\max}/2$.

Note that the Kullback-Leibler divergence between two Poisson distributions with rates λ and λ' is

$$\mathrm{D_{KL}}(\lambda \,\|\, \lambda') = \lambda' - \lambda + \lambda \log \frac{\lambda}{\lambda'}$$

$$\leq \lambda' - \lambda + \lambda \left(\frac{\lambda}{\lambda'} - 1 \right)$$

$$= \frac{(\lambda - \lambda')^2}{\lambda'}.$$

Therefore, for any $\theta \in \Theta$,

$$\mathrm{D_{KL}}(P_\theta \,\|\, Q) \leq rk\ell p \frac{\delta^2}{\lambda_{\max}/2}$$

$$= m\ell p \frac{2\delta^2}{\lambda_{\max}}.$$

Take

$$\delta = \sqrt{\frac{\lambda_{\max}}{32\ell p}} \leq \frac{\lambda_{\max}}{2}.$$

Then

$$p_e \geq \frac{1}{2} - \frac{8 \log 2}{m},$$

and half the separation between points indexed by Θ is at least

$$\frac{\sqrt{m\ell}\delta}{2} = \frac{1}{8\sqrt{2}} \sqrt{\frac{m\lambda_{\max}}{p}} = \frac{1}{8\sqrt{2}} \sqrt{\frac{r}{p}} \sigma_1.$$

$\square$

3.2 Second lower bound

The previous theorem relies on the fact that the observations are conditionally Poisson. The next result, which provides the second part of a matching lower bound to (5), does not depend on the conditional distribution of the observations, and instead shows a fundamental limit in inferring missing matrix entries.

Theorem 4. *Take again $m = rk$, $n = r\ell$. Set $\sigma_2^2 = k\lambda_{\max}^2$. Let*

$$S = \left\{ M \in [0, \lambda_{\max}]^{m \times n} : \mathrm{rank}(M) \le r, \sqrt{\max_i \sum_j M_{ij}^2} + \sqrt{\max_j \sum_i M_{ij}^2} \le 2\sigma_2 \right\}.$$

Suppose $p \ge \frac{1}{2(k \wedge \ell)} = \frac{r}{2(m \wedge n)}$. Then, under a Bernoulli sampling model with probability p (with any conditional distribution on the observations),

$$\inf_{\widehat{M}} \sup_{M \in S_2} \mathbf{E} \|\widehat{M} - M\|_F^2 \ge \frac{r\sigma_2^2}{8} \max \left\{ \frac{1}{2} \left\lfloor \frac{1}{2p} \right\rfloor, 1 - p \right\}$$

$$\ge \frac{1}{64} \frac{1-p}{p} r\sigma_2^2.$$

Proof. Again, we assume that $k \ge \ell$. We first prove the lower bound with the first item in the maximum. For notational simplicity, we can assume that $1/2p$ is an integer. Furthermore, we can assume that $\ell = 1/2p$, since decreasing the number of columns does not increase risk.

We consider a set of matrices $\{M_\theta\}_{\theta \in \{0,1\}^m}$ with the same structure as in the proof of Theorem 3, but now, we take $\lambda_0 = 0$, $\lambda_1 = \lambda_{\max}$.

Assouad's lemma (see, e.g., [34] or [33]) gives a lower bound on the Bayes risk of an estimator $\widehat{M}$ for a uniform prior on $\{0,1\}^m$:

$$R(\widehat{M}) := \frac{1}{2^m} \sum_{\theta \in \{0,1\}^m} \mathbf{E}_\theta \|\widehat{M} - M_\theta\|_F^2$$

$$\ge \frac{1}{2} \sum_{i=1}^m \frac{\ell \lambda_{\max}^2}{4} \inf_\phi \left(\mathbf{P}_\theta(\phi \ne \theta_i) + \mathbf{P}_{\theta^i}(\phi \ne 1 - \theta_i) \right),$$

where $\phi \colon \mathbf{R}^{m \times n} \to \{0,1\}$ is a test, and, for $\theta \in \{0,1\}^m$, θ^i denotes the element of $\{0,1\}^m$ that is equal to θ except in the ith position.

Denote by P_θ^i the marginal distribution of the ith row of a matrix with distribution P_θ. The minimal testing risk which appears in the sum is equal to the L_1 norm of the minimum of the densities of P_θ^i and $P_{\theta^i}^i$, which measures how much the distributions overlap. Thus we have

$$\inf_\phi \left(\mathbf{P}_\theta(\phi \ne \theta_i) + \mathbf{P}_{\theta^i}(\phi \ne 1 - \theta_i) \right) = \|P_\theta \wedge P_{\theta^i}\|_1$$

$$\ge (1 - p)^\ell$$

$$\ge 1 - \ell p$$

$$= \frac{1}{2},$$

where the first inequality is due to the fact that, with probability $(1 - p)^\ell$, no entry from the ith row of M is observed.

Then

$$R(\widehat{M}) \ge \frac{m\ell \lambda_{\max}^2}{16} = \frac{r\sigma_2^2}{32p}.$$

The result follows from the fact that minimax risk always exceeds Bayes risk.

A simple modification with $\ell = 1$ yields the result for the second term in the maximum. $\qquad \square$

We note here that we could also use a similar argument as in the proof of Theorem 3 to get a high-probability lower bound on error (with a somewhat worse constant). For example, setting $Q = \frac{P_{(0,\dots,0)} + P_{(1,\dots,1)}}{2}$, it is easily verified that the resulting Kullback-Leibler divergences $\mathrm{D}_{\mathrm{KL}}(P_\theta \| Q)$ can be upper bounded by a sum of coordinate-wise total variation distance.

17

3.3   When do the upper and lower bounds match?

Within multiplicative constants, the lower bounds of Theorems 3 and 4 match the approximate upper bound of (5). We must therefore consider when the approximation in (5) is accurate.

Finding technical conditions that guarantee matching rates is not something we think likely to be very instructive at this point, especially since a different proof technique could potentially change the logarithmic term in $A(M, p, \epsilon)$. However, we think it is helpful to look at the matrices involved in the proofs of Theorems 3 and 4. Note that when the bounds match for those particular matrices, the minimax error rate bounds are tight for the matrix classes considered in those proofs.

For these matrices (assuming that $m \geq n$),

$$\widetilde{\sigma} \approx \sqrt{\frac{m}{r}}(\sqrt{\lambda_{\max}} + \sqrt{1-p}\lambda_{\max}).$$

For this term to dominate (6), we must have

$$\widetilde{\sigma} \gtrsim \frac{\lambda_{\max} \vee c\log m}{\sqrt{p}}\sqrt{\log m}.$$

For example, if $\lambda_{\max} \geq \log m$, it would suffice to take

$$p \gtrsim \frac{r\log m}{m},$$

which is a standard condition in noiseless matrix completion. If $\lambda_{\max} \leq \log m$, it would suffice to take

$$p \gtrsim \frac{r\log^3 m}{m\lambda_{\max}^2}.$$

4   Conclusion and future work

In this paper, we have derived an upper bound in Frobenius norm error for an estimator for Poisson matrix completion, and we have derived a minimax lower bound that matches this upper bound (within a universal constant) for many classes of nonnegative rate matrices. We have also derived similar upper bounds for error in two types of multinomial matrix denoising problems. The estimators we use are computationally tractable, and require significantly fewer assumptions on the underlying matrix than previous results in the literature. Significantly, we impose no lower bounds on the entries of the underlying matrix. This is crucial in many applications (such as topic modelling) where zero or very small means can be relatively common.

Because we have found upper and lower error bounds in Frobenius norm, the only theoretical improvement remaining for this model and error metric in general classes of matrices is to try to relax the conditions under which the bounds match (although, as we have seen, they are not too restrictive now). This could potentially come about by reducing the logarithmic term in (6) and/or by finding a logarithmic term to add to the minimax lower bounds. One could also further examine how to obtain better error rates for more restrictive classes of matrices, such as incoherent matrices.

It would also be interesting to extend the results presented here to matrices that are not exactly low-rank, but are instead "approximately low-rank"; for example, we could consider matrices which are contained in Schatten balls (which, for $q \in [0,1]$, are sets of matrices for which $\sum_i \sigma_i^q \leq R$, where $\{\sigma_i\}$ is the set of singular values). As mentioned previously, Cao and Xie [15] used the Schatten 1-norm ($q = 1$, or nuclear norm ball); Negahban and Wainwright [1] also examined these classes of matrices.

Another avenue of research would be to examine structured Poisson or multinomial estimation under different, more statistically motivated error metrics. Maximum likelihood methods seem more suitable here than least-squares, but analysis of maximum likelihood estimators has proved difficult for the reasons outlined in Section 1.5. It is not clear what kind of structure would be relevant in a different error metric. Low-rank structure seems to work well with a least-squares error framework, but there is *a priori* not much reason to think that it would work similarly well for another metric; for example, the Bhattacharyya distance between Poisson distributions, is proportional to the (squared) ℓ_2 distance between the *square roots* of the rates, but the element-wise square root of a low-rank matrix is not, in general, low rank. Thus, this approach may not immediately bear much fruit. However, an analysis of matrix estimation under alternative error metrics remains an important area for future research.

Acknowledgments

References

[1] S. Negahban and M. Wainwright, "Restricted strong convexity and weighted matrix completion: Optimal bounds with noise," *J. Mach. Learn. Res.*, vol. 13, pp. 1665–1697, May 2012.

[2] V. Koltchinskii, K. Lounici, and A. Tsybakov, "Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion," *Ann. Stat.*, vol. 39, no. 5, pp. 2302–2329, Oct. 2011.

[3] E. Candes and Y. Plan, "Matrix completion with noise," *Proc. IEEE*, vol. 98, no. 6, pp. 925–936, Jun. 2010.

[4] X. Zhou, C. Yang, H. Zhao, and W. Yu, "Low-rank modeling and its applications in image analysis," *ACM Comput. Surv.*, vol. 47, no. 2, 36:1–36:33, Dec. 2014.

[5] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. ACM SIGIR Conf. (SIGIR)*, Berkeley, CA, Aug. 1999.

[6] D. Lee and S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, Oct. 1999.

[7] J. Canny, "GaP: A factor model for discrete data," in *Proc. ACM SIGIR Conf. (SIGIR)*, Sheffield, United Kingdom, Jul. 2004.

[8] A. Cemgil, "Bayesian inference for nonnegative matrix factorisation models," *Comput. Intell. Neurosci.*, 2009.

[9] A. Mnih and R. Salakhutdinov, "Probabilistic matrix factorization," in *Proc. Conf. Neural Inf. Process. Syst. (NIPS)*, Vancouver, Canada, Dec. 2007.

[10] H. Ma, C. Liu, I. King, and M. Lyu, "Probabilistic factor models for web site recommendation," in *Proc. ACM SIGIR Conf. (SIGIR)*, Beijing, China, Jul. 2011.

[11] P. Gopalan, J. Hofman, and D. Blei, "Scalable recommendation with hierarchical poisson factorization," in *Proc. Conf. Uncert. in Artif. Intell. (UAI)*, Amsterdam, Netherlands, Jul. 2015.

[12] E. Candes and T. Tao, "The Dantzig selector: Statistical estimation when p is much larger than n," *Ann. Stat.*, vol. 35, no. 6, pp. 2313–2351, 2007.

[13] R. Keshavan, A. Montanari, and S. Oh, "Matrix completion from noisy entries," *J. Mach. Learn. Res.*, vol. 11, pp. 2057–2078, 2010.

[14] R. Mazumder, T. Hastie, and R. Tibshirani, "Spectral regularization algorithms for learning large incomplete matrices," *J. Mach. Learn. Res.*, vol. 11, pp. 2287–2322, 2010.

[15] Y. Cao and Y. Xie, "Poisson matrix recovery and completion," *IEEE Trans. Signal Process.*, vol. 64, no. 6, pp. 1609–1620, Mar. 2016.

[16] A. Soni, S. Jain, J. Haupt, and S. Gonella, "Noisy matrix completion under sparse factor models," *IEEE Trans. Inf. Theory*, vol. 62, no. 6, pp. 3636–3661, Jun. 2016.

[17] A. Soni and J. Haupt, "Estimation error guarantees for Poisson denoising with sparse and structured dictionary models," in *Proc. IEEE Int. Symp. on Inf. Theory (ISIT)*, Honolulu, HI, Jun. 2014.

[18] O. Klopp, "Matrix completion by singular value thresholding: Sharp bounds," *Electron. J. Stat.*, vol. 9, pp. 2348–2369, 2015.

[19] ——, "Noisy low-rank matrix completion with general sampling distribution," *Bernoulli*, vol. 20, pp. 282–303, 1 2014.

[20] S. Gunasekar, A. Banerjee, and J. Ghosh, "Unified view of matrix completion under general structural constraints," in *Proc. Conf. Neural Inf. Process. Syst. (NIPS)*, Montreal, Canada, Dec. 2015.

[21] S. Gaïffas and G. Lecué, "Sharp oracle inequalities for high-dimensional matrix prediction," *IEEE Trans. Inf. Theory*, vol. 57, no. 10, pp. 6942–6957, Oct. 2011.

[22] O. Klopp, "Rank penalized estimators for high-dimensional matrices," *Electron. J. Stat.*, vol. 5, pp. 1161–1183, 2011.

[23] S. Chatterjee, "Matrix estimation by universal singular value thresholding," *Ann. Stat.*, vol. 43, pp. 177–214, 1 2015.

[24] S. Gaïffas and O. Klopp, "High dimensional matrix estimation with unknown variance of the noise," *Stat. Sin.*, vol. 27, pp. 115–145, 2017.

[25] J. Lafond, "Low rank matrix completion with exponential family noise," in *Proc. Conf. Learn. Theory (COLT)*, Paris, France, Jul. 2015.

[26] S. Gunasekar, P. Ravikumar, and J. Ghosh, "Exponential family matrix completion under structural constraints," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Beijing, China, Jun. 2014.

[27] G. Raskutti, M. Wainwright, and B. Yu, "Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls," *IEEE Trans. Inf. Theory*, vol. 57, no. 10, pp. 6976–6994, Oct. 2011.

[28] A. Bandeira and R. van Handel, "Sharp nonasymptotic bounds on the norm of random matrices with independent entries," *Ann. Probab.*, vol. 44, no. 4, pp. 2479–2506, Jul. 2016.

[29] Q. Huang, S. Kakade, W. Kong, and G. Valiant, "Recovering structured probability matrices," in *Proc. Innovations Theor. Comput. Sci. (ITCS)*, Cambridge, MA, Jan. 2018.

[30] R. Latała, R. van Handel, and P. Youssef, "The dimension-free structure of nonhomogeneous random matrices," *Invent. Math.*, vol. 214, pp. 1031–1080, 3 Dec. 2018.

[31] D. Pollard, "MiniEmpirical," Available at `http://www.stat.yale.edu/~pollard/`, 2016.

[32] J. Tropp, "An introduction to matrix concentration inequalities," *Found. Trends Mach. Learn.*, vol. 8, pp. 1–230, 1-2 May 2015.

[33] C. Huber, "Lower bounds for function estimation," in *Festschrift for Lucien Le Cam*, D. Pollard, E. Torgersen, and G. L. Yang, Eds., New York: Springer, 1997, ch. 15, pp. 245–258.

[34] B. Yu, "Assouad, Fano, and Le Cam," in *Festschrift for Lucien Le Cam*, D. Pollard, E. Torgersen, and G. L. Yang, Eds., New York: Springer, 1997, ch. 29, pp. 423–435.