Near-Optimal Entrywise Anomaly Detection for Low-Rank Matrices with Sub-Exponential Noise

Vivek F. Farias 1  Andrew A. Li 2  Tianyi Peng 3

Abstract
We study the problem of identifying anomalies in a low-rank matrix observed with sub-exponential noise, motivated by applications in retail and inventory management. State of the art approaches to anomaly detection in low-rank matrices apparently fall short, since they require that non-anomalous entries be observed with vanishingly small noise (which is not the case in our problem, and indeed in many applications). So motivated, we propose a conceptually simple entrywise approach to anomaly detection in low-rank matrices. Our approach accommodates a general class of probabilistic anomaly models. We extend recent work on entrywise error guarantees for matrix completion, establishing such guarantees for sub-exponential matrices, where in addition to missing entries, a fraction of entries are corrupted by (also unknown) anomaly model. Viewing the anomaly detection as a classification task, to the best of our knowledge, we are the first to achieve the min-max optimal detection rate (up to log factors). Using data from a massive consumer goods retailer, we show that our approach provides significant improvements over incumbent approaches to anomaly detection.

1. Introduction
Consider the problem of identifying anomalies in a low-rank matrix: specifically, let $M^*$ be some low-rank matrix, and let $X = M^* + E + A$, where $E$ is a noise matrix with independent, mean-zero entries, and where $A$ is a sparse matrix of anomalies. We observe only $X$, and only on some subset of matrix entries $\Omega$. The anomaly detection problem concerns identifying the support of $A$ simply from this observation.

One of the most popular approaches to solving this problem is the following convex optimization formulation, referred to as ‘stable principal component pursuit’ (stable PCP) (Zhou et al., 2010),

$$
\min_{\hat{M}, \hat{A}} \|\hat{M}\|_* + \lambda_2 \|\hat{A}\|_1 + \lambda_1 \|P_\Omega(X - \hat{M} - \hat{A})\|_F^2, \tag{1}
$$

where $\lambda_1$ and $\lambda_2$ are regularization parameters. The three matrix norms in the objective are meant to promote, from left to right, low-rankredness in $\hat{M}$, sparsity in $\hat{A}$, and fit to $X$ on the observed entries $\Omega$. Upon solving problem (1), $\hat{A}$ can be used to estimate the support of $A$. Now in the absence of anomalies, this optimization problem (after removing the $\hat{A}$ terms) is in essence optimal under a variety of assumptions on the distributions of $E$ and $\Omega$. In contrast, the available results for anomaly detection are weaker. Perhaps most limiting, results that guarantee recovery of $M^*$ require that the ‘average’ noise $\|E\|_F/n$ vanishes, where $n$ is the size of the matrix. In this setting, noise in observing any individual matrix entry in $\Omega$ grows negligibly small in large matrices. This is limiting:

1. $X$ is typically noisy: In the practical problem that motivates this work, $M^* + E$ can be viewed as a matrix of centered Poisson entries with mean $M^*$. Clearly then, $E\|E\|_F$ will scale with the size of the matrix, so theoretical guarantees for extant anomaly detection approaches do not apply.

2. Even ignoring this theoretical limitation, we will see that in the setting where $X$ is noisy (such as in our motivating application), the optimization approach above can perform quite poorly.

---

1 Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
2 Tepper School of Business, Carnegie Mellon University, Pittsburgh, PA 15213, USA
3 Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

Correspondence to: Vivek F. Farias <vivekf@mit.edu>, Andrew A. Li <aali1@cmu.edu>, Tianyi Peng <tianyi@mit.edu>.
1.1. Overview of Main Contributions

Against the above backdrop, we make the following contributions to the problem of anomaly detection in matrices:

1. **An optimal algorithm**: We develop a new algorithm for low-rank matrices with sub-exponential noise, and prove that our approach, for the first time, achieves the min-max optimal anomaly detection rate (up to logarithmic terms) under a broad class of probabilistic anomaly models (Theorem 1 and Proposition 1).

2. **Entrywise guarantee for sub-exponential matrices**: As part of our approach, we prove a new recovery guarantee of independent interest (Theorem 2) for the matrix completion problem. This result is unique in applying to sub-exponential (vs. the usual sub-gaussian) noise, and bounding the entrywise (vs. an aggregate) error.

3. **Applications**: Our work is motivated by a crucial inventory management problem (‘phantom inventory’) that costs the retail industry up to 4% in annual revenue. We observe that this inventory problem can be viewed as one of detecting anomalies in a low-rank Poisson matrix. The latter is the matrix one obtains by viewing sales data in matrix-form with rows corresponding to store locations, columns corresponding to products, and entries corresponding to observed sales over some time. On large-scale data (thousands of stores, thousands of products) we find that our approach achieves 13% and 19% higher accuracy (measured via the usual area under ROC curve) than convex optimization approaches on synthetic and real data, respectively.

1.2. Related Literature

There are three ongoing streams of work to which the present paper contributes. The first, naturally, is in anomaly detection for matrices. The majority of existing work has focused on a formulation called robust principal component analysis (robust PCA) (Candès et al., 2011; Chandrasekaran et al., 2011). Most relevant to our problem are approaches for noisy robust PCA (Zhou et al., 2010; Agarwal et al., 2012; Wong & Lee, 2017; Klopp et al., 2017; Chen et al., 2020b). See Table 1 for a summary of existing results. Note that any hope of identifying the anomalies $A$ would require, at the very least, that $\|M - M^*\|_F = o(n)$. Thus, with respect to the noisy problem we are studying, in which $\|E\| = \Omega(n)$, existing results are insufficient. In contrast, our algorithm not only improves upon the recovery of $M^*$ to sufficiently allow for recovery of $A$, it also provides an additional guarantee on entrywise recovery: $\|\hat{M} - M^*\|_{\text{max}}$. All our guarantees are min-max optimal, and beyond the recovery of $M^*$, to the best of our knowledge, we are also the first paper to analyze the anomaly detection rate as a formal classification problem. Finally, concurrent with this paper, (Chen et al., 2020b) provide a more-refined analysis of Eq. (1) and achieve a recovery guarantee for $M^*$ similar to ours. However, they require a zero-mean assumption on the anomalies, i.e. $\mathbb{E}(A) = 0$, which is not representative of many applications, such as the ones motivating this work.

The second body of work concerns statistical inference in matrix completion (Abbe et al., 2017; Chen et al., 2019; Ma et al., 2019; Chen et al., 2020a). Our contribution here is an entrywise error bound for matrix completion under sub-exponential noise. This result substantially improves upon previous results for sub-exponential matrices, all of which bound an aggregate error measure (Lafond, 2015; Sambasivan & Haupt, 2018; Cao & Xie, 2015; McRae & Davenport, 2019). Our analysis builds on a recent framework introduced in (Abbe et al., 2017) for sub-gaussian noise, and requires both a considerably more fine-tuned computation, and drawing from a recent result for Poisson matrix completion from (McRae & Davenport, 2019).

Finally, with respect to our motivating application: the phantom inventory problem is well-studied in the field of Operations Management (Raman et al., 2001; DeHoratius & Raman, 2008; Nachtmann et al., 2010; Fan et al., 2014; Chen & Mersereau, 2015; Wang et al., 2016). Existing algorithmic solutions (Kök & Shang, 2007; DeHoratius et al., 2008) have focused on adapting inventory management policies to

| $(\text{Zhou et al., } 2010)$ | $n \|E\|_F$ | $-$ |
| $(\text{Wong & Lee, } 2017)$ | $\sqrt{n} \|E\|_F$ | $-$ |
| $(\text{Klopp et al., } 2017)$ | $\sqrt{\log n} \|E\|_F$ | $-$ |
| **This paper** | $\frac{\sqrt{\log n}}{\sqrt{n}} \|E\|_F$ | $\frac{\sqrt{\log n}}{n \sqrt{n}} \|E\|_F$ |

Table 1: Comparison of our results with existing work under proper hyper-parameters. The reported quantities are the scalings of upper bounds on the error of $\|\hat{M} - M^*\|$, for two matrix norms, with respect to matrix size $n$ and noise $E$.

---

1As pointed out in (Chen et al., 2020b), this assumption is likely a fundamental limitation of Eq. (1).
Our goal is to infer $B$ we assume that entries are served independently with probability $p$ and are the consequence of so-called shelf-execution errors and typically result in a censoring of sales so that for our motivating problem $\text{Anom}(\alpha^*, \lambda)$ is perhaps best viewed as a censored Poisson($\lambda$) random variable. Our results will allow for a broad family of distributions for anomalies, which we describe momentarily.

**Assumptions on $M^*$:** Let $M^* = U\Sigma V^T$, be the SVD of $M^*$, where $\Sigma \in \mathbb{R}^{r \times r}$ is a diagonal matrix with singular values $\sigma_1^* \geq \sigma_2^* \geq \ldots \geq \sigma_r^*$ ($\kappa = \sigma_1^*/\sigma_r^*$); and $U \in \mathbb{R}^{m \times r}, V \in \mathbb{R}^{m \times r}$ are two matrices that hold the left and right-singular vectors. We make the following assumptions:

- (Boundedness): $\|M^*\|_{\max} + 1 \leq L$.
- (Incoherence): $\|U\|_{\infty} + \|V\|_{\infty} \leq \sqrt{\frac{\mu}{n+m}}$.
- (Sparsity): $\sqrt{\frac{\mu}{n+m}} \geq C_1 \frac{\log^{1.5}(m) \mu r^2 L^2}{\|M^*\|_{\max} \sqrt{m}}$ for some known constant $C_1$.

Our guarantees will be parameterized by $\mu, L, r$ and $\kappa$. These assumptions and parameters for $M^*$ are similar to those in the existing matrix completion literature (Abbe et al., 2017; Ma et al., 2019).

**Assumptions on $\text{Anom}(\cdot, \cdot)$:** We make the following assumptions:

- (Sub-exponential): $\text{Anom}(\alpha^*, M^*_{ij})$ is sub-exponential: $\|\text{Anom}(\alpha^*, M^*_{ij})\|_{\psi_1} \leq L$.
- (Lipschitz): For any $M \in \mathbb{R}_+, \alpha \in \Gamma$ and all $k \in \mathbb{N}$, $\mathbb{P} \left( \text{Anom}(\alpha, M) = k \right)$ is $K$-Lipschitz in $(\alpha, M)$.
- (Mean Decomposition): For any $M \in \mathbb{R}_+, \alpha \in \Gamma$, we have $\mathbb{E} \left( \text{Anom}(\alpha, M) \right) = g(\alpha) M$ for some $g : \mathbb{R}^d \to \mathbb{R}$ where $g(\alpha)$ is $K$-Lipschitz in $\alpha$.

We pause to discuss these assumptions on anomalies. To begin, we assume a probabilistic anomaly model characterized by finite unknown parameters $\alpha$. This generally applies to many applications. The probabilistic and Lipschitz properties enable the identification of $\alpha$ and then the measure of the anomaly detection rate. We also assume a mean-decomposition condition: $\mathbb{E}(A) \propto M^*$. Note that this is less restrictive than the zero-mean assumption $\mathbb{E}(A) = 0$ sometimes found in the literature. It is also well justified from the known mechanism for anomalies in the phantom inventory problem such as $\text{Anom}(\alpha, M^*_{ij}) = \text{Poisson}(\alpha M^*_{ij})$. (DeHoratius et al., 2008).

In contrast to this probabilistic model, one could consider an adversarial anomaly model. Note that the adversarial model that allows for arbitrary anomalies requires in essence exact observations of $M^*$ for non-anomalous observations (Candès et al., 2011), which is not consistent with our highly noisy setup.
2.1. Performance Metrics

Let \( A^\pi(X_\Omega) \) be some estimator of \( B \). Given \( X_\Omega \),
we define the true positive rate for this estimator,
\( \text{TPR}_\pi(X_\Omega) \) as the ratio of the expected number of true
positives under the algorithm and the expected number of anomalies
given \( X_\Omega \). We similarly define the false positive rate,
\( \text{FPR}_\pi(X_\Omega) \). More formally, let \( f^*_\pi \)
be the conditional probability that the \((i, j)\)-th entry
is not anomalous, given \( X \), i.e.
\[
f^*_\pi := \Pr(B_{ij} = 0 \mid X).
\]

Then, some algebraic manipulation establishes that
\[
\text{TPR}_\pi(X_\Omega) = \frac{\sum_{(i, j) \in \Omega} \Pr(A^\pi_{ij}(X_\Omega) = 1)(1 - f^*_\pi)}{\sum_{(i, j) \in \Omega}(1 - f^*_\pi)}
\]
\[
\text{FPR}_\pi(X_\Omega) = \frac{\sum_{(i, j) \in \Omega} \Pr(A^\pi_{ij}(X_\Omega) = 1)f^*_\pi}{\sum_{(i, j) \in \Omega}f^*_\pi}.
\]

Our goal will be to maximize TPR for some bound on
FPR. In establishing the quality of our algorithm we
will compare, for a given constraint on FPR, the TPR
achieved under our algorithm to that achieved under the
(clairvoyant) optimal estimator. We will show that
in large matrices this gap grows negligibly small at a
min-max optimal rate.

3. Algorithm and Results

We are now prepared to state our approach to the
anomaly detection problem formulated above. Our
algorithm, which we refer to as the entrywise (EW)
algorithm, leverages an entrywise matrix completion
guarantee for sub-exponential noise that we will
describe shortly. Besides the observed data \( X_\Omega \),
the only other input into the EW algorithm is a target FPR
which we denote as \( \gamma \). The full algorithm is stated in
Algorithm 1 below:

The goal of the EW algorithm is to maximize the
TPR subject to a FPR below the input target value of
\( \gamma \). Our main result is the following guarantee, which
states that (a) the ‘hard’ constraint on the FPR is sat-
sified with high probability, and (b) the TPR is within
an additive regret of a certain unachievable policy we
use as a proxy for the best achievable policy. Specif-
cally, for any \( \gamma \in (0, 1] \), let \( \pi^*(\gamma) \) denote the optimal
policy when \( M^*, p^*_\Lambda \), and \( \alpha^* \) are known (this policy
is described later in this section). One can verify that,
for any \( \gamma \), \( X_\Omega \) and policy \( \pi \), \( \text{TPR}_{\pi^*(\gamma)}(X_\Omega) \geq \text{TPR}_{\pi}(X_\Omega) \)
if \( \text{FPR}_\pi(X_\Omega) \leq \gamma \). Note that the only additional as-
sumptions we require, beyond those stated in Section
2, are the set of regularity conditions (RC) stated in
the following section.

**Algorithm 1** Entrywise (EW) Algorithm \( \pi^{EW}(\gamma) \)

**Input:** \( X_\Omega, \gamma \in (0, 1] \)

1. Set \( M = \frac{nm}{\| \Omega \|} \text{SVD}(X_\Omega) \). Here \( \text{SVD}(X_\Omega) := \text{arg min}_{\text{rank}(M) \leq r} \| M - X' \|_F \),
   where \( X' \) is obtained from \( X_\Omega \) by setting unobserved entries to 0.
2. Estimate \( \hat{\theta} = (\hat{p}_A, \hat{a}) \) based on the moment-
   matching estimator in Eq. (2).
3. Estimate a confidence interval \([f^1_{ij}, f^R_{ij}]\) for \( f^*_{ij} \)
   with each \((i, j) \in \Omega \) according to Eq. (3).
4. Let \( \{t^E_{ij}\} \) be an optimal solution to the following
   optimization problem:
   \[
   \begin{align*}
   p^{EW} := & \max_{\{0 \leq t_{ij} \leq 1, (i, j) \in \Omega\}} \sum_{(i, j) \in \Omega} t_{ij} \\
   \text{subject to} & \sum_{(i, j) \in \Omega} t_{ij} f^R_{ij} \leq \gamma \sum_{(i, j) \in \Omega} t_{ij} f^L_{ij}
   \end{align*}
   \]
5. For every \((i, j) \in \Omega \), generate \( A_{ij} \sim \text{Ber}(t^E_{ij}) \) in-
dependently.

**Output:** \( A_\Omega \)

**Theorem 1.** Assume that the regularity conditions
(RC) hold. With probability \( 1 - O(\frac{1}{nm}) \), for any \( 0 < \gamma \leq 1 \),
\[
\begin{align*}
\text{FPR}_{\pi^{EW}(\gamma)}(X_\Omega) & \leq \gamma, \\
\text{TPR}_{\pi^E(\gamma)}(X_\Omega) & \geq \text{TPR}_{\pi^*(\gamma)}(X_\Omega) - C(K + L)^3L^3K^4\mu^r \log^{1.5}(m) \sqrt{p_\Lambda \gamma}.
\end{align*}
\]

To parse this result, consider that in a typical applica-
tion, we can expect the problem parameters to fall
in the following scaling regime: \( K, L, n, r, \mu = O(1),
\) \( p_\Lambda, p^*_\alpha, \gamma = \Omega(1), \) and \( m/n = \Theta(1) \).
For this regime, the regret is \( O(n^{-1/2} \log^{1.5} n) \),
which is in fact optimal up to logarithmic factors. To be precise, we fix a
particular value of \( \gamma \) for which the following proposition
states that, for any \( n \), there exists a family of anomaly
models \( \mathcal{M}_n \) for which no algorithm can achieve a regret
on TPR lower than \( O(n^{-1/2}) \) across all models within
the family (we will explicitly construct this family in
the proof in Section 4.4). To allow for direct compari-
son to Theorem 1, let \( \Pi_\gamma \) denote the set of all policies \( \pi \)
such that
\[
\Pr_{X|M^*}(\text{FPR}_\pi(X) \leq \gamma) \geq 1 - C/n^2 \quad \text{for all } M^* \in \mathcal{M}_n.
\]

**Proposition 1.** For any algorithm \( \pi \in \Pi_\gamma \), there ex-
ists \( M^* \in \mathcal{M}_n \) such that
\[
\text{E}_{X|M^*}[\text{TPR}_{\pi^*(\gamma)}(X) - \text{TPR}_\pi(X)] \geq C/\sqrt{n}.
\]
This shows that our algorithm is optimal for the TPR up to logarithmic factors.

Algorithmic Novelty: Before proceeding to the sketches of these results, it is worth discussing the novelty of our algorithm. The vast majority of existing algorithms all seek to decompose $X$ into its three components ($M, A, E$) by solving a single optimization problem: $\min f(M) + g(A) + h(E)$, where $f, g, h$ are carefully chosen penalty functions, e.g., Eq. (1). In contrast to these approaches, our algorithm uses two separate procedures. The first procedure (Step 1) is effectively a de-noising and completion routine which, rather unintuitively, makes no attempt to identify the positions of anomalies, but is able to estimate $M$ with a guarantee on entrywise accuracy, but only up to an unknown affine scaling. This entrywise guarantee (a fundamentally new result not previously exploited by the aforementioned optimization algorithms) enables the second procedure (Steps 2–5), which leverages the underlying probabilistic structure to estimate the affine scaling and perform entrywise inference, yielding the first sharp statements about the optimality of the anomaly detection rate.

4. Algorithm Details and Proof Sketches

In this section, we motivate the steps of Algorithm 1 and sketch the proofs of our main results. Beginning with Theorem 1, and mirroring the algorithm itself, the following description is given in four parts: (i) an entrywise guarantee for $\hat{M}$; (ii) a moment matching estimator for $\hat{\theta}$; (iii) a confidence interval for $f^{ij}_{\hat{M}}$; (iv) an analysis of the optimization problem $\mathcal{P}^{EW}$.

4.1. Step 1: Entrywise Guarantee for $\hat{M}$

Our algorithm is initiated with a de-noising of $X_{\Omega}$. To ease notation, let $\theta = (p_A, \alpha), \theta^* = (p^*_{A}, \alpha^*)$, and denote $e(\theta) := p_A g(\alpha) + (1 - p_A)$. This latter function is chosen so that, as follows from a quick calculation, $E(X) = e(\theta^*) M^*$. While the SVD-based de-noising algorithm used here is standard, the key result that drives rest of the algorithm and analysis is the following new entrywise error bound, which may be of independent interest:

**Theorem 2.** Let

$$\hat{M} = \frac{nm}{|\Omega|} \text{SVD}(X_{\Omega})_r.$$

Then with probability $1 - O(\frac{1}{nm})$,

$$\|\hat{M} - e(\theta^*) M^*\|_{\infty} \leq C \kappa^4 \rho L \sqrt{\log(m)} \sqrt{\frac{\log(m)}{p_{\text{OM}}}}.$$

Our result can be viewed as the first entrywise guarantee result for Poisson matrix completion.\(^3\) The proof sketch is provided in the Appendix. As a comparison, consider the recent results for aggregated error on matrix completion with Poisson noise (McRae & Davenport, 2019). Under the proper hyperparameters, their results based on SVD provide the following Frobenius norm bound: $\|\hat{M} - M^*\|_F \lesssim n^{1/2}$. In contrast, our entrywise guarantees provide that $\|\hat{M} - M^*\|_{\infty} \lesssim n^{-1/2} \log^{1/2} n$. Therefore, our results show that the SVD approach not only provides aggregated error guarantee but also yields a much stronger result: the entrywise error guarantee. Furthermore, the entrywise error is evenly distributed among all entries up to a logarithmic factor.

The entrywise guarantee is the key that opens up optimal anomaly detection. In particular, this enables us in the next steps to infer both the parameters $\theta^*$ and the posterior probabilities of anomalies at each entry.

4.2. Step 2: Moment Matching Estimator

Step 1 yields an (entrywise) accurate estimator $\hat{M}$ of $M^*$, but only up to some linear scaling that depends on the unknown anomaly model parameters $\theta^*$. Now in Step 2, we are able to use $\hat{M}$ to estimate that unknown scaling $e(\theta^*)$, along with $\theta^*$ itself, via a generalized moment of the cumulative distribution function at sufficiently many values for identifiability. In particular, for any $t \in \mathbb{N}$, let $g_t(\theta, M)$ be the proportion of entries of $X_{\Omega}$ expected to be at most $t$ with the model specified by $\theta$ and $M$:

$$g_t(\theta, M) := \mathbb{E}(|X_{ij} \leq t, (i, j) \in \Omega|)/\mathbb{E}(|\Omega|).$$

Given that $M^* \approx \hat{M}/e(\theta^*)$, we choose $\hat{\theta}$ to be the minimizer of the following function which seeks to match a set of $T$ empirical moments to their expectations as closely as possible (in $\ell^2$ distance),

$$\hat{\theta} := \arg\min_{\theta \in \Theta} \sum_{t=0}^{T-1} \left( g_t(\hat{\theta}, \hat{M}/e(\theta)) - \frac{|X_{ij} \leq t, (i, j) \in \Omega|}{|\Omega|} \right)^2$$

(2)

where $T$ is a large enough constant for identifiability (usually $T = d + 1$ for $\theta \in \mathbb{R}^{d+1}$) and $\Theta$ is chosen such

\(^3\)In fact, the proof also holds valid for sub-exponential noise.
Near-Optimal Anomaly Detection for Matrices with Sub-Exponential Noise

that \( \theta^* \in \Theta \) and \( p_A \) is bounded from 1 by a constant for \( \theta = (p_A, \alpha) \in \Theta \).

Let \( F = (F_0, F_1, \ldots, F_{T-1}) : \Theta \to \mathbb{R}^T \) be defined as \( F_t(\theta) = g_t(\theta, M^t/\epsilon(\theta^*)/\epsilon(\theta)) \). We could expect that \( F_t(\hat{\theta}) \approx F_t(\theta^*) \) by solving \( \hat{\theta} \) from Eq. (2). In fact, we have the following result.

**Lemma 1.** With probability \( 1 - O(\frac{1}{nm}) \),

\[
\left\| F(\hat{\theta}) - F(\theta^*) \right\| \leq C(K + L)\kappa^4 \mu r L \sqrt{\frac{\log(m)}{p_0 m}}.
\]

To establish that \( \hat{\theta} \approx \theta^* \) from \( F(\hat{\theta}) \approx F(\theta^*) \), additional regularity conditions are required. Let \( \delta' = \kappa^4 \mu r L \frac{\log m}{p_0 m} \) be the entrywise bound of \( \|M - \epsilon(\theta^*)M^*\|_{\text{max}} \). We now formally state the regularity conditions that we require:

(\text{RC}) Regularity Conditions on \( F(\theta) \):

- \( F : \Theta \to \mathbb{R}^T \) is continuously differentiable and injective.
- Let \( \delta = \delta'(K + L) \log m \). We require \( B_{\delta}(\theta^*) \subset \Theta \), where \( B_{\delta}(\theta^*) = \{ \theta : \|\theta - \theta^*\| \leq r \} \).
- For any \( \theta \in B_{\delta}(\theta^*) \), \( \|J_F(\theta) - J_F(\theta^*)\|_2 \leq \frac{C}{\delta} \|\theta - \theta^*\| \), where \( J \) is the Jacobian matrix.
- \( \|J_F(\theta^*)^{-1}\|_2 \leq C \).

These conditions are among the typical set of conditions for methods involving generalized moments and are well justified in typical applications (Newey & McFadden, 1994; Imbens et al., 1995; Hall, 2005; Hansen, 1992). The following lemma establishes that our moment matching estimator is able to accurately estimate \( \theta^* \).

**Lemma 2.** Assuming the above regularity conditions (RC) on \( F(\theta) \), with probability \( 1 - O(\frac{1}{nm}) \),

\[
\left\| \hat{\theta} - \theta^* \right\| \leq C(K + L)\kappa^4 \mu r L \sqrt{\frac{\log m}{p_0 m}}.
\]

Extending to general noise models: To extend Algorithm 1 to general sub-exponential noise, all steps hold the same except that the estimator for \( \hat{\theta} \) in the Step 2 needs to be changed. For observation \( X \) with continuous values, one can use MLE estimator to solve \( \hat{\theta} = \arg\max_\theta P(X|\theta, M^*) \) by plugging in \( M^* \approx M/e(\theta) \). For integer-value \( X \) beyond Poisson noise, we still use moment matching estimator (incorporate negative values in Eq. (2) if needed). In both cases, the results in this paper will still hold with slightly different regularity conditions for the identification of parameters.

4.3. Steps 3–5: Confidence Intervals and the Optimization Problem \( \mathcal{P}_{\text{EW}} \):

Next, we use \( \hat{M} \) and \( \hat{\theta} \) as plug-in estimators to optimize TPR under the FPR constraint. Let

\[
\hat{x}_{ij} := [\hat{p}_A P_{\text{Anom}}(X_{ij}|\hat{\alpha}, \hat{M}_{ij}/e(\hat{\theta}))],
\]

\[
\hat{y}_{ij} := [(1 - \hat{p}_A) P_{\text{Poisson}}(X_{ij}|\hat{M}_{ij}/e(\hat{\theta}))],
\]

where \( [x] \) denotes \( x \) ‘truncated’ to its nearest value in \([0, 1]\), i.e. \( [x] = \max(\min(x, 1), 0) \). Then we can estimate a confidence interval \([f^L_{ij}, f^R_{ij}]\) for each conditional non-anomaly probability \( f^*_{ij} = P(B_{ij} = 0 | X) \) using what effectively amounts to a plug-in estimator based on \( \hat{x}_{ij}, \hat{y}_{ij} \). That is the content of the following result:

**Lemma 3.** Let

\[
\delta = (K + L)^3\kappa^4 \mu r L^2 \sqrt{\frac{\log m}{p_0 m}}.
\]

There exists a (known) constant \( C_1 \) such that, if

\[
f^L_{ij} := \left[ \frac{\hat{y}_{ij} - C_1 \delta}{\hat{x}_{ij} + \hat{y}_{ij}} \right] \quad \text{and} \quad f^R_{ij} := \left[ \frac{\hat{y}_{ij} + C_1 \delta}{\hat{x}_{ij} + \hat{y}_{ij}} \right],
\]

then with probability \( 1 - O(\frac{1}{nm}) \), for every \((i, j) \in \Omega\), we have

\[
f^L_{ij} \leq f^*_{ij} \leq f^R_{ij} + \epsilon_{ij} \quad \text{and} \quad f^R_{ij} - \epsilon_{ij} \leq f^*_{ij} \leq f^R_{ij},
\]

where \( \epsilon_{ij} = \min(4C_1 \delta / (\hat{x}_{ij}^* + \hat{y}_{ij}^*), 1) \).

The final two steps involve solving \( \mathcal{P}_{\text{EW}} \). To motivate its particular form, consider the ‘ideal’ anomaly detection algorithm if the \( f^*_{ij} \)’s were known. Intuitively, one should identify anomalies at entries with the smallest values of \( f^*_{ij} \). This leads to the following idealized algorithm, which we will call \( \pi^*(\gamma) \):

1. Let \( \{t^*_{ij}\} \) be an optimal solution to the following optimization problem.

\[
\mathcal{P}^* : \max_{\{0 \leq t_{ij} \leq 1\} \in \Omega} \sum_{(i,j) \in \Omega} t_{ij}
\]

subject to \( \sum_{(i,j) \in \Omega} t_{ij} f^*_{ij} \leq \gamma \sum_{(i,j) \in \Omega} f^*_{ij} \)

2. For every \((i, j) \in \Omega\), generate \( A_{ij} \sim \text{Ber}(t^*_{ij}) \) independently.

The following claim establishes the optimality of \( \pi^*(\gamma) \).
Claim 1. For any $\pi, \gamma$, and $X_\Omega$, if $\text{FPR}_\pi(X_\Omega) \leq \gamma$, then $\text{TPR}_\pi(X_\Omega) \leq \text{TPR}_{\pi^*}(X_\Omega)$.

Now notice that $\mathcal{P}^{\text{EW}}$ is obtained from $\mathcal{P}^*$ by replacing $f_{ij}^L$ with the confidence interval estimators $f_{ij}^C$ and $f_{ij}^R$ defined in the previous step. Intuitively, we could expect that $\mathcal{P}^{\text{EW}} \approx \mathcal{P}^*$, and therefore the algorithm $\pi^{\text{EW}}$ should achieve the desired performance. In fact, $\text{FPR}_{\pi^{\text{EW}}}(X) \leq \gamma$ holds immediately because $f_{ij}^L \leq f_{ij}^* \leq f_{ij}^R$ and so $\{t_{ij}^{\text{EW}}\}$ is a feasible solution of $\mathcal{P}^*$. The guarantee for $\text{TPR}_{\pi^{\text{EW}}}(X)$ can be established based on a fine-tuned analysis of Lemma 3. See the Appendix for the formal proof.

4.4. Minimax Lower Bound

In this final subsection, we provide a sketch for showing Proposition 1 based on the hypothesis testing argument. We consider the following special model: let $p_A = 1$ and $p_A^* = \frac{1}{2}$, and when $B_{ij} = 1$, $X_{ij} = 0$. We refer to this in notational form as $X \sim H(M^*)$.

We construct $\mathcal{M}_n = \{M^b \in \mathbb{R}^{n \times n}, b \in \{0, 1\}^{n/2}\}$ as follows. Fix a constant $C$. Consider $b \in \{0, 1\}^{n/2}$. For any $i \in [n/2], j \in [n]$, if $b_i = 0$, $M_{2i,j}^b = 1$ and $M_{2i+1,j}^b = 1 - \frac{C}{\sqrt{n}}$; if $b_i = 1$, $M_{2i,j}^b = 1 - \frac{C}{\sqrt{n}}$ and $M_{2i+1,j}^b = 1$. Let $M^b$ be drawn uniformly from $\mathcal{M}_n$ and $X \sim H(M^b)$. In order to achieve high TPR, one needs to identify the correct $b$ given $X$. By our construction, the error probability for distinguishing $M_1^b$ and $M_2^b$ is large when $b_1$ is 1-bit different from $b_2$. This provides a lower bound on identifying $b$ and leads to a $O(1/\sqrt{n})$ regret on TPR. The formal proof can be found in the Appendix. One can also verify that Theorem 1 guarantees $O(\log^{1.5} n/\sqrt{n})$ regret for every $M^b \in \mathcal{M}_n$. This shows that our algorithm is optimal for the TPR up to logarithmic factors.

5. Experiments

To evaluate the empirical performance of the EW algorithm, we first consider a synthetic setting where we compare the performances of our EW algorithm with various state-of-the-art approaches. We then measure performance on real-world data from a large retailer.

5.1. Synthetic Data

We generated an ensemble of 1000 matrices $M^*$ of size $n = m = 100$. The varying parameters of the ensemble include (i) $r$: the rank of the matrix; (ii) $M^* = \frac{1}{nm} \sum_{ij} M^*_{ij}$: the average value of all entries; (iii) $p_0$: the probability of an entry being observed; (iv) $p_{\lambda}^*$: the probability of an entry where an anomaly occurs; and (v) $\alpha^*$: the anomaly parameter. When an anomaly occurs, $E(\text{Anom}(\alpha^*, M)) = \alpha^* M$.

The parameters were sampled uniformly: $r \in [1, 10], M^* \in [1, 10], p_0 \in [0.5, 1], p_{\lambda}^* \in [0, 0.3]$ and $\alpha^* \in [0, 1]$. Each instance was generated in two steps: (i) Generate $M^*$: for a given choice of $r$ and entry-wise mean $M^*$, we set $M^* = kUV^T$. $U, V \in \mathbb{R}^{n \times r}$ are random with independent Gamma(1, 2) entries and $k$ is picked so that $M^* = \frac{1}{\sqrt{n}} \sum_{ij} M^*_{ij}$. This is a typical way of generating $M^*$ with rank $r$ and non-negative entries (Cemgil, 2008). (ii) Observation: If $(i, j)$ is observed, then with probability $1 - p_{\lambda}^*, X_{ij} \sim \text{Poisson}(M_{ij});$ otherwise, $X_{ij} \sim \text{Poisson}(\text{Exp}(\alpha^*) M_{ij})$. Here Exp($\alpha^*$) models the occurring time of the anomalous event.

We compared our EW algorithm with three existing algorithms: (i) Stable-PCP (Zhou et al., 2010; Chen et al., 2020b), (ii) Robust Matrix Completion (RMC) (Klopp et al., 2017), and (iii) Direct Robust Matrix Factorization (DRMF) (Xiong et al., 2011). These three algorithms all recover the matrices by decomposing $X = \hat{M} + \hat{A} + \hat{E}$ and minimizing $f(\hat{M}) + \lambda_1 g(\hat{A}) + \lambda_2 h(\hat{E})$ where $f, g, h$ are penalty functions with Lagrange multipliers $\lambda_1, \lambda_2$. For all algorithms, we tuned Lagrange multipliers corresponding to rank using knowledge of the true rank. In order to generate ROC curves and compute AUCs, we varied $\gamma$ in our EW algorithm. For three existing optimization algorithms, we do this by varying the Lagrange multipliers.

The results are summarized in Table 2 and Figure 1. Table 2 reports AUC, $||\hat{M} - M||_F,$ and $||\hat{M} - M||_{\max}$ averaged over 1000 instances for four algorithms ($\hat{M}$ of EW is obtained after recovering from the estimated scaling). The results show that EW achieves an AUC close to $\pi^*$ (the ideal algorithm that knows $M^*$ and the anomaly model), confirming Theorem 1. For all considered metrics, the results also demonstrate that EW outperforms other algorithms significantly. Figure 1 (Left) shows that the above phenomenon holds
Near-Optimal Anomaly Detection for Matrices with Sub-Exponential Noise

Figure 1: **Synthetic data.** (Left) Scatter plot showing AUC of ideal algorithm vs. that of EW (blue points, above 45-degree line); and AUC of Stable PCP vs EW (green, mostly below 45 degree line). (Right) ROC curve in a representative setting with \( n = m = 100, r = 3, \bar{M}^* = 5, p_O = 0.8, p_A^* = 0.04, \alpha^* = 0.2 \).

Figure 2: **Real data.** (Left) An ensemble similar to the synthetic data. (Right) ROC curve in a representative setting of \( p_A^* = 0.04 \) and \( \alpha^* = 0.2 \).

uniformly over the ensemble.\(^4\) Figure 1 (Right) shows the explicit ROC curve for a representative setting.

### 5.2. **Real Data**

We collected data, from a retailer, consisting of weekly sales of \( m = 290 \) products across \( n = 2481 \) stores with \( p_O \sim 0.14 \) and mean value 2.64. Since there is no ground-truth for anomalies, we viewed the collected data as the underlying matrix \( M^* \), and then introduced noise and artificial anomalies. Specifically, we generated \( X \) as in the synthetic data (with fixed \( M^* \)), introducing anomalies by deliberately perturbing a fraction \( p_A^* \) of entries and thinning the resulted sales at rate \( \alpha^* \). In particular, for each sample, \( p_A^* \in [0, 0.3], \alpha^* \in [0, 1] \) were uniformly drawn. We generated an ensemble of 1000 such perturbed matrices.

Figure 2 reports the results. We see similar relative merits as in the synthetic experiments: EW achieves an AUC close to that of an algorithm that knows \( M^* \) and \( \alpha^* \) whereas Stable PCP is consistently worse than EW. In particular, the average AUC of \( \pi^* \) is 0.733, the average AUC of \( \pi^{EW} \) is 0.672, whereas the average AUC of Stable-PCP is 0.566. Right of the Figure 2 shows an ROC curve for a representative setting of \( p_A^* = 0.04 \) and \( \alpha^* = 0.2 \)\(^5\) where we see the absolute performance. Our results also confirm the EW’s ability to recover \( M^* \) for real-world data and simulated anomalies. We left the experiments on real anomalies as the future work. More details about experiments

\(^4\)We show results vs. Stable-PCP, but the same holds true for the other two existing algorithms in the experiments.

\(^5\)The parameters are chosen to fit the reported loss caused by the phantom inventory (Gruen et al., 2002).
can be found in the Appendix.

**Scalability:** EW is also much faster than compared algorithms, since our main computational cost is a typical matrix completion procedure (our linear program step can be solved via a sort). Concretely, we can expect to solve a $70000 \times 10000$ matrix with $10^7$ observed entries within minutes (Yao & Kwok, 2018).

6. Conclusion

We proposed a simple statistical model for anomaly detection in low-rank matrices that is motivated by the phantom inventory problem in retail. We proved a new entrywise bound for matrix completion with sub-exponential noise, and used this to motivate a simple policy for the anomaly detection problem. We proved matching upper and lower bounds on the anomaly detection rate of our algorithm, and demonstrated in experiments that our approach provides substantial improvements over existing approaches.

References

Abbe, E., Fan, J., Wang, K., and Zhong, Y. Entrywise eigenvector analysis of random matrices with low expected rank. *arXiv preprint arXiv:1709.09565*, 2017.

Agarwal, A., Negahban, S., Wainwright, M. J., et al. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *The Annals of Statistics*, 40(2):1171–1197, 2012.

Bernstein, S. The theory of probabilities, 1946.

Candès, E. J., Li, X., Ma, Y., and Wright, J. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011.

Cao, Y. and Xie, Y. Poisson matrix recovery and completion. *IEEE Transactions on Signal Processing*, 64(6):1609–1620, 2015.

Cemgil, A. T. Bayesian inference for nonnegative matrix factorisation models. *Computational intelligence and neuroscience*, 2009, 2008.

Chandrasekaran, V., Sanghavi, S., Parrilo, P. A., and Willsky, A. S. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

Chen, J., Liu, D., and Li, X. Nonconvex rectangular matrix completion via gradient descent without $l_{2,\infty}$ regularization. *IEEE Transactions on Information Theory*, 2020a.

Chen, L. and Mersereau, A. J. Analytics for operational visibility in the retail store: The cases of censored demand and inventory record inaccuracy. In *Retail supply chain management*, pp. 79–112. Springer, 2015.

Chen, Y., Fan, J., Ma, C., and Yan, Y. Inference and uncertainty quantification for noisy matrix completion. *arXiv preprint arXiv:1906.04159*, 2019.

Chen, Y., Fan, J., Ma, C., and Yan, Y. Bridging convex and nonconvex optimization in robust pca: Noise, outliers, and missing data. *arXiv preprint arXiv:2001.05484*, 2020b.

Conrad, S. Sales data and the estimation of demand. *Journal of the Operational Research Society*, 27(1):123–127, 1976.

DeHoratius, N. and Raman, A. Inventory record inaccuracy: An empirical analysis. *Management science*, 54(4):627–641, 2008.

DeHoratius, N., Mersereau, A. J., and Schrage, L. Retail inventory management when records are inaccurate. *Manufacturing & Service Operations Management*, 10(2):257–277, 2008.

Fan, T.-J., Chang, X.-Y., Gu, C.-H., Yi, J.-J., and Deng, S. Benefits of rfid technology for reducing inventory shrinkage. *International Journal of Production Economics*, 147:659–665, 2014.

Gross, D. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.

Gruen, T. W., Corsten, D. S., and Bharadwaj, S. Retail out-of-stocks: A worldwide examination of extent, causes and consumer responses. *Grocery Manufacturers of America*, Washington, DC, 2002.

Hall, A. R. *Generalized method of moments*. Oxford university press, 2005.

Hansen, L. P. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pp. 1029–1054, 1982.

Imbens, G. W., Johnson, P., and Spady, R. H. Information theoretic approaches to inference in moment condition models. Technical report, National Bureau of Economic Research, 1995.

Klopp, O., Lounici, K., and Tsybakov, A. B. Robust matrix completion. *Probability Theory and Related Fields*, 169(1-2):523–564, 2017.
Near-Optimal Anomaly Detection for Matrices with Sub-Exponential Noise

Kök, A. G. and Shang, K. H. Inspection and replenishment policies for systems with inventory record inaccuracy. *Manufacturing & service operations management*, 9(2):185–205, 2007.

Lafond, J. Low rank matrix completion with exponential family noise. In *Conference on Learning Theory*, pp. 1224–1243, 2015.

Ma, C., Wang, K., Chi, Y., and Chen, Y. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. *Foundations of Computational Mathematics*, pp. 1–182, 2019.

Ma, S. and Aybat, N. S. Efficient optimization algorithms for robust principal component analysis and its variants. *Proceedings of the IEEE*, 106(8):1411–1426, 2018.

Mazumder, R., Hastie, T., and Tibshirani, R. Spectral regularization algorithms for learning large incomplete matrices. *Journal of machine learning research*, 11(Aug):2287–2322, 2010.

McRae, A. D. and Davenport, M. A. Low-rank matrix completion and denoising under poisson noise. *arXiv preprint arXiv:1907.05325*, 2019.

Nachtmann, H., Waller, M. A., and Rieske, D. W. The impact of point-of-sale data inaccuracy and inventory record data errors. *Journal of Business Logistics*, 31(1):149–158, 2010.

Newey, K. and McFadden, D. Large sample estimation and hypothesis. *Handbook of Econometrics, IV, Edited by RF Engle and DL McFadden*, pp. 2112–2245, 1994.

Raman, A., DeHoratius, N., and Ton, Z. Execution: The missing link in retail operations. *California Management Review*, 43(3):136–152, 2001.

Sambasivan, A. V. and Haupt, J. D. Minimax lower bounds for noisy matrix completion under sparse factor models. *IEEE Transactions on Information Theory*, 64(5):3274–3285, 2018.

Shi, J., Katehakis, M. N., Melamed, B., and Xia, Y. Production-inventory systems with lost sales and compound poisson demands. *Operations Research*, 62(5):1048–1063, 2014.

Tropp, J. A. et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

Wang, F., Fang, X., Chen, X., and Li, X. Impact of inventory inaccuracies on products with inventory-dependent demand. *International Journal of Production Economics*, 177:118–130, 2016.

Wong, R. K. and Lee, T. C. Matrix completion with noisy entries and outliers. *The Journal of Machine Learning Research*, 18(1):5404–5428, 2017.

Xiong, L., Chen, X., and Schneider, J. Direct robust matrix factorization for anomaly detection. In *Proceedings of the 2011 IEEE 11th International Conference on Data Mining*, pp. 844–853. IEEE, 2011.

Yao, Q. and Kwok, J. T. Accelerated and inexact soft-impute for large-scale matrix and tensor completion. *IEEE Transactions on Knowledge and Data Engineering*, 31(9):1665–1679, 2018.

Zhou, Z., Li, X., Wright, J., Candès, E., and Ma, Y. Stable principal component pursuit. In *2010 IEEE international symposium on information theory*, pp. 1518–1522. IEEE, 2010.
Appendices. Proof of Statements

A. Entry-wise Bound and Proof of Theorem 2

In this section, we will prove the Theorem 2 based on recent results on entry-wise analysis for random matrices (Abbe et al., 2017) and matrix completion with Poisson observation (McRae & Davenport, 2019). The proof idea can be viewed as a generalization from Gaussian noise in the Theorem 3.4 (Abbe et al., 2017) to subexponential noise. In particular, we will proceed the proof in two steps: (i) consider the symmetric scenario where \( M^* \), noises, and anomalies have symmetries; (ii) generalize the results to the asymmetric scenario.

A.1. Symmetric Case

Consider a symmetric scenario. Let \( \tilde{M}^* \in \mathbb{R}_+^{n \times n} \) be a symmetric matrix. For \( 1 \leq i \leq j \leq n \), let

\[
\begin{align*}
\tilde{X}_{ij} &\sim \text{Poisson}(M_{ij}^*) \quad \text{with prob. } (1 - p^*)_{ij} \rho_0 \\
\tilde{X}_{ij} &\sim \text{Anom}(a^*, \tilde{M}_{ij}^*) \quad \text{with prob. } p^* \rho_0 \\
\tilde{X}_{ij} &= 0 \quad \text{with prob. } 1 - \rho_0.
\end{align*}
\]

(4)

Let \( \tilde{X}_{ij} = \tilde{X}_{ij} \) for \( 1 \leq i \leq j \leq n \). Let \( t = g(a^*)p^*\rho_0 + (1 - p^*)\rho_0 \). It is easy to verify \( \mathbb{E}(\tilde{X}/t) = \tilde{M}^* \). Furthermore, suppose \( \max(\tilde{M}^*_i + 1, \|\text{Anom}(a^*, \tilde{M}_{ij}^*)\|_{\psi_1} ) \leq L \) for \( (i, j) \in [n] \times [n] \).

Denote the eigenvalues of \( \tilde{M}^* \) by \( \lambda_1^* \geq \lambda_2^* \geq \ldots \geq \lambda_n^* \) with their associated eigenvectors by \( \{\tilde{u}_j^*\}_{j=1}^n \). Denote the eigenvalues of \( \tilde{X} \) by \( \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \) with their associated eigenvectors by \( \{\tilde{u}_j\}_{j=1}^n \).

Suppose \( r \) is an integer such that \( 1 \leq r < n \). Assume \( \tilde{M}^* \) satisfies \( \lambda_1^* \geq \lambda_2^* \geq \ldots \geq \lambda_r^* \geq 0 \) and \( \lambda_{r+1}^* \leq 0 \). Let \( \tilde{U} = (u_1^*, u_2^*, \ldots, u_r^*) \in \mathbb{R}^{n \times r} \). We aim to show that \( \tilde{U} \) is a good estimation of \( \tilde{U}^* \) in the entry-wise manner under some proper rotation. In particular, let \( \tilde{H} := \tilde{U}^T \tilde{M}^* \in \mathbb{R}^{r \times r} \). Suppose the SVD decomposition of \( \tilde{H} \) is \( \tilde{H} = U\Sigma V^T \) where \( U, V \in \mathbb{R}^{r \times r} \) are orthonormal matrices and \( \Sigma \in \mathbb{R}^{r \times r} \) is a diagonal matrix. The matrix sign function of \( \tilde{H} \) is denoted by \( \text{sgn}(\tilde{H}) := UV^T \). In fact, \( \text{sgn}(\tilde{H}) = \arg \min_{\tilde{O}} \|\tilde{U} \tilde{O} - \tilde{U}^*\|_F \) subject to \( \tilde{O} \tilde{O}^T = I \).

Let \( \Delta^* := t\lambda_r^*, \kappa := \frac{\lambda_1^*}{\lambda_r^*} \). We rephrase the Theorem 2.1 in (Abbe et al., 2017) for the above scenario and rewrite it as the following lemma.

**Lemma 6 (Theorem 2.1 (Abbe et al., 2017)).** Suppose \( \gamma \in \mathbb{R}_{\geq 0} \). Let \( \phi(x) : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0} \) be a continuous and non-decreasing function with \( \phi(0) = 0 \) and \( \phi(x)/x \) non-increasing in \( \mathbb{R}_{\geq 0} \). Let \( \delta_0, \delta_1 \in (0, 1) \). With the above quantities, consider the following four assumptions:

**A.1.** \[ \|\tilde{M}^*\|_{2,\infty} \leq \gamma \Delta^*. \]

**A.2.** For any \( m \in [n] \), the entries in the \( m \)th row and column of \( \tilde{X} \) are independent with others.

**A.3.** \[ \mathbb{P}
\left(
\|\tilde{X} - t\tilde{M}^*\|_2 \leq \gamma \Delta^* \right) \geq 1 - \delta_0. \]

**A.4.** For any \( m \in [n] \) and any \( \tilde{W} \in \mathbb{R}^{n \times r} \),

\[
\mathbb{P}
\left(
\|\tilde{X} - t\tilde{M}^*\|_m, W\|_2 \leq \Delta^* \|W\|_{2,\infty} \phi \left( \frac{\|W\|_F}{\sqrt{\mathbb{E}(\|W\|_F^2)}} \right) \right) \geq 1 - \delta_1/n.
\]

If \( 32\kappa \max(\gamma, \phi(\gamma)) \leq 1 \), under above Assumptions A1–A4, with probability \( 1 - \delta_0 - 2\delta_1 \), the followings hold,

\[
\|\tilde{U}\|_{2,\infty} \lesssim (\kappa + \phi(1)) \|\tilde{U}^*\|_{2,\infty} + \gamma \|t\tilde{M}^*\|_{2,\infty} / \Delta^*.
\]

\[
\|\tilde{U}\text{sgn}(\tilde{H}) - \tilde{U}^*\|_{2,\infty} \lesssim (\kappa + \phi(1)(\gamma + \phi(\gamma))) \|\tilde{U}^*\|_{2,\infty} + \gamma \|t\tilde{M}^*\|_{2,\infty} / \Delta^*.
\]

To obtain useful results from Lemma 6, one need to find proper \( \gamma \) and \( \phi(x) \) and show that the Assumptions A1–A4 hold. We define \( \tilde{\gamma} \) and \( \tilde{\phi}(x) \) as the proper form for \( \gamma \) and \( \phi(x) \) respectively in the following.

---

6See (Gross, 2011) for more details about the matrix sign function.
Definition 2. Let $\tilde{\gamma} := \sqrt{\frac{n^2}{2} L}, \tilde{\phi}(x) := \frac{\sqrt{2}}{\sqrt{n} L} \log(2n^3 r) x$.

Under $\tilde{\gamma}$ and $\tilde{\phi}(x)$, we will show that Assumption A3 holds based on Lemma 11, Assumption A4 holds based on Lemma 13. Note that Assumption A2 naturally holds since each element of $X$ is independent of each other. Assumption A1 holds due to that $\|t\bar{M}^*\|_{2,\infty} = \max_i \sqrt{\sum_j t^2 \bar{M}_{ij}^2} \leq t\sqrt{n} L \leq \tilde{\gamma} \Delta^*$.

To show that Assumption A3 holds, we introduce a result in (McRae & Davenport, 2019) that helps to control the operator norm of $\bar{X} - t\bar{M}^*$.

**Lemma 7** (Lemma 4 in (McRae & Davenport, 2019)). Let $Y$ be a random $n_1 \times n_2$ matrix whose entries are independent and centered, and suppose that for some $v, t_0 > 0$, we have, for all $t_0 \geq t_0$, $\mathbb{P}(|Y| \geq t_1) \leq 2e^{-t_1/v}$. Let $\epsilon \in (0, 1/2)$, and let $K = \max\{t_0, \sqrt{\log(2mn/\epsilon)}\}$. Then,

$$
\mathbb{P}\left(\|Y\|_2 \geq 2\sigma + \frac{cv}{\sqrt{n_1 n_2}} + t_1\right) \leq \max(n_1, n_2) \exp(-t_1^2/(C_0(2K)^2)) + \epsilon,
$$

where $C_0$ is a constant and $\sigma = \max_i \sqrt{\sum_j \mathbb{E} \left(\frac{Y^2_{ij}}{\sigma^2}\right)} + \max_j \sqrt{\sum_i \mathbb{E} \left(\frac{Y^2_{ij}}{\sigma^2}\right)}$.

In order to use Lemma 7, we show that every entry of $\bar{X} - t\bar{M}^*$ is a sub-exponential random variable based on Lemmas 8 to 10.

**Lemma 8.** Let $Y \sim \text{Poisson}(\lambda)$. Then $\|Y\|_{\psi_1} \leq 4\lambda + 1$.

**Proof.** Note that for any $t_1 > 0$,

$$
\mathbb{E} \left(e^{\|Y\|_{\psi_1}} \right) = \mathbb{E} \left(e^{Y/t_1} \right) = \sum_{k=0}^{\infty} \frac{e^{k/t_1}}{k!} \lambda^k e^{-\lambda} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(e^{1/t_1} \lambda)^k}{k!} = e^{-\lambda} e^{e^{1/t_1} \lambda} = e^{\lambda(e^{1/t_1} - 1)}.
$$

Note that $1/(4\lambda + 1) \leq 1$, hence $e^{1/(4\lambda + 1) - 1} = \frac{1}{4\lambda + 1} e^s \leq \frac{1}{4\lambda + 1} e$ where $s \in [0, 1/(4\lambda + 1)]$ by Taylor expansion. Therefore

$$
\mathbb{E} \left(e^{\|Y\|_{\psi_1}} \right) \leq e^{\frac{\lambda(e^{1/t_1} - 1)}{4\lambda + 1}} = e^{\frac{1}{4}} \approx 1.973 < 2.
$$

By the definition of $\|\cdot\|_{\psi_1}$, we have $\|Y\|_{\psi_1} \leq 4\lambda + 1$.

**Lemma 9.** Let $Y_1, Y_2, \ldots, Y_q$ be q subexponential random variables with $\|Y_i\|_{\psi_1} \leq L_{\max}$. Let $c \in \{1, 2, \ldots, q\}$ be a random variable. Then $\|Y_c\|_{\psi_1} \leq L_{\max}$.

**Proof.** This is because $\mathbb{E} \left(e^{\|Y\|_{\psi_1}} \right) = \sum_{i=1}^{\infty} \mathbb{P}(c = i) \mathbb{E} \left(e^{\|Y_i\|_{\psi_1}} \right) \leq \sum_{i=1}^{q} 2\mathbb{P}(c = i) = 2$.

**Lemma 10.** For any $(i, j) \in [n] \times [n]$, $\|\bar{X}_{ij} - t\bar{M}_{ij}^*\|_{\psi_1} \leq 6L$.

**Proof.** Note that $\|\text{Poisson}(\bar{M}_{ij}^*)\|_{\psi_1} \leq 4L$ by Lemma 8 and $\|\text{Anom}(\alpha, \bar{M}_{ij}^*)\| \leq L$ by the definition of $L$. We have $\|\bar{X}_{ij}\|_{\psi_1} \leq 4L$ by Eq. (4) and Lemma 9. Then, by the triangle inequality, $\|\bar{X}_{ij} - t\bar{M}_{ij}^*\|_{\psi_1} \leq \|\bar{X}_{ij}\|_{\psi_1} + \|t\bar{M}_{ij}^*\|_{\psi_1} \leq 4L + 2L = 6L$.

Next we show that Assumption A3 holds.

**Lemma 11.** Suppose $p_0 \geq \frac{\log^3 n}{n}$. $\mathbb{P}(\|\bar{X} - t\bar{M}^*\|_2 \leq C\tilde{\gamma} \Delta^*) \geq 1 - \frac{1}{n^2}$ for some constant $C$.

**Proof.** Denote $Y \in \mathbb{R}^{n \times n}$ by

$$
Y_{ij} = \begin{cases} 
2(\bar{X}_{ij} - t\bar{M}_{ij}^*) & i < j \\
(\bar{X}_{ij} - t\bar{M}_{ij}^*) & i = j \\
0 & i > j
\end{cases}
$$
Note that $\|Y_{ij}\|_{\psi_1} \leq 2\|\tilde{X}_{ij} - tM^*_{ij}\|_{\psi_1} \leq 12L$ by Lemma 10. By the property of subexponential random variable, for all $t' \geq 0$, $P(|Y_{ij}| \geq t') \leq 2\exp(-t'/C_1L)$ where $C_1$ is a constant. By the construction of $Y$, we also have

$$E(Y_{ij}) = 0 \quad \text{and} \quad E(Y_{ij}^2) \leq 2E[\tilde{X}_{ij}^2] \leq C_2pOL^2$$

for some constant $C_2$.

Consider applying Lemma 7 to $X$ with $n_1 = n_2 = n$. Let $\epsilon = \frac{1}{3\pi^2}$. Then $K = C_1L\log(4n^2)$. Take $t' = \sqrt{C_0\log n} + \log 2$. Then $\max(n_1, n_2)\exp(-t^2/(C_0(2K)^2)) + \epsilon = \frac{1}{\pi^2}$. Furthermore, by Eq. (5) and $npo \geq \log^3(n)$, one can verify that

$$2\sigma + \frac{n\epsilon}{\sqrt{n_1n_2}} + t' \leq C_3\sqrt{npOL}$$

for some constant $C_3$.

Therefore, $\|Y\|_2 \leq C_3\sqrt{npOL}$ with probability $1 - \frac{1}{\pi^2}$. Note that $\tilde{X} - tM^* = (Y + Y^T)/2$. Hence, with probability $1 - \frac{1}{\pi^2}$, $\|X - tM^*\|_2 \leq (\|Y\|_2 + \|Y^T\|_2)/2 \leq C_3\sqrt{npOL} \leq C_3\gamma^*$. \hfill \Box

Next, we will show that Assumption A4 holds based on the matrix Bernstein’s inequality to control the tail bound of sum of sub-exponential random variables.

**Lemma 12** (Matrix Bernstein’s inequality). Given $n$ independent random $m_1 \times m_2$ matrices $X_1, X_2, \ldots, X_n$ with $E[X_i] = 0$. Let

$$V \triangleq \max\left(\left\|\sum_{i=1}^n E[X_iX_i^T]\right\|, \left\|\sum_{i=1}^n E[X_i^TX_i]\right\|\right).$$

Suppose $\|X_i\|_{\psi_1} \leq L$ for $i \in [n]$. Then,

$$\|X_1 + X_2 + \ldots + X_n\| \lesssim \sqrt{V \log(n(m_1 + m_2)) + L \log(n(m_1 + m_2)) \log(n)}$$

with probability $1 - O(n^{-c})$ for any constant $c$.

**Proof.** Let $Y_i = X_i1 \{\|X_i\| \leq B\}$ be the truncated version of $X_i$. We have,

$$\|E(Y_i)\| \leq \left\|\int X_i1 \{\|X_i\| > B\} df(X_i)\right\| \leq (i) \int \|X_i\|1 \{\|X_i\| > B\} df(X_i) \leq BP(\|X_i\| > B) + \int_B^\infty P(\|X_i\| > t)dt \leq (ii) Be^{-B/C_L} + CLe^{-B/C_L}$$

where (i) is due to the convexity of $\|\cdot\|$ and (ii) is due to the subexponential property of $\|X_i\|$ and $C$ is a constant.

Meanwhile, we have

$$\left\|\sum_{i=1}^n E((Y_i - E(Y_i))(Y_i - E(Y_i))^T)\right\| \leq \left\|\sum_{i=1}^n E(Y_iY_i^T) - E(Y_i)E(Y_i)^T\right\| \leq (i) \left\|\sum_{i=1}^n E(Y_iY_i^T)\right\| \leq \left\|\sum_{i=1}^n E(X_iX_i^T) - E(X_iX_i^T1 \{\|X_i\| > B\})\right\| \leq (ii) \left\|\sum_{i=1}^n E(X_iX_i^T)\right\| \leq V$$
where (i) is due to the positive-semidefinite property of $E(Y_i) E(Y_i)^T$ and $E(Y_i Y_i^T) - E(Y_i) E(Y_i)^T$, (ii) is due to the positive-semidefinite property of $E(X_i X_i^T 1 \{\|X_i\| > B\})$ and $E(Y_i Y_i^T)$. Similarly, $\|\sum_{i=1}^n E((Y_i - E(Y_i))^T (Y_i - E(Y_i)))\| \leq V$.

Then, by Theorem 6.1.1 (Tropp et al., 2015), we have

$$\mathbb{P}\left(\left\|\sum_{i=1}^n (Y_i - E(Y_i))\right\| \geq t\right) \leq 2 \exp\left(\frac{-t^2/2}{V + 2B t/3}\right).$$

Then, with probability $1 - O(n^{-c})$ for some constant $c$,

$$\left\|\sum_{i=1}^n (Y_i - E(Y_i))\right\| \lesssim \sqrt{V \log(n(m_1 + m_2))} + B \log(n(m_1 + m_2)).$$

Take $B = L \log(n) C'$ for a proper constant $C'$, by Eq. (8), we have

$$\left\|\sum_{i=1}^n Y_i\right\| \lesssim \sqrt{V \log(n)} + L \log^2(n) + nL \log(n) O(n^{-C'/C})$$

$$\lesssim \sqrt{V \log(n(m_1 + m_2))} + L \log(n(m_1 + m_2)) \log(n).$$

By the union bound on the event $\|X_i\| \leq B$ for all $i$, we can conclude that, with probability $1 - O(n^{-c'})$ for some constant $c'$,

$$\left\|\sum_{i=1}^n X_i\right\| \lesssim \sqrt{V \log(n(m_1 + m_2))} + L \log(n(m_1 + m_2)) \log(n).$$

Consider the Assumption A4.

**Lemma 13.** For any $m \in [n]$ and any $W \in \mathbb{R}^{n \times r}$, the following holds

$$\mathbb{P}\left(\left\|\bar{X} - tM^\ast\right\|_2 \leq C\Delta^\ast \|W\|_{2,\infty} \phi\left(\frac{\|W\|_{fp}}{\sqrt{n} \|W\|_{2,\infty}}\right) \right) \geq 1 - O(n^{-3})$$

where $C$ is a constant.

**Proof.** Let $Y_j = X_{ij} - tM^\ast_{ij}$ and $Z_j = Y_j W_{j\cdot} \in \mathbb{R}^{1 \times r}$. Note that

$$\left\|\left(\bar{X} - tM^\ast\right)_{m\cdot} W\right\|_2 = \left\|\sum_{j=1}^n Z_j\right\|_2.$$

We aim to invoke the Lemma 12 for $Z_1, Z_2, \ldots, Z_n$. Note that $E(Z_j) = 0$ since $E(Y_j) = 0$ and $Z_j$ are independent since $Y_j$ are independent. Also, for the subexponential norm, we have

$$\|Z_j\|_{\psi_1} \leq \|Y_j\|_{\psi_1} \|W_{j\cdot}\|_2$$

$$\leq \|Y_j\|_{\psi_1} \|W\|_{2,\infty}$$

$$\lesssim L \|W\|_{2,\infty}.$$
where the last inequality is due to Lemma 10. Then, one can check
\[
\left\| \sum_{j=1}^{n} \mathbb{E} \left( Z_j^T Z_j \right) \right\| \leq \sum_{j=1}^{n} \left\| \mathbb{E} \left( Z_j^T Z_j \right) \right\| \\
\leq \sum_{j=1}^{n} \mathbb{E} \left( Y_j^2 \right) \left\| W_{j*} \right\|_2^2 \\
\leq \sum_{j=1}^{n} L^2 p_0 \left\| W_{j*} \right\|_2^2 \\
\leq L^2 p_0 \left\| W \right\|_F^2.
\]
Similarly, one can show that
\[
\left\| \sum_{j=1}^{n} \mathbb{E} \left( Z_j^T Z_j \right) \right\| \leq L^2 p_0 \left\| W \right\|_F^2.
\]
Then, with probability \(1 - O(n^{-3})\),
\[
\left\| (\bar{X} - tM^*)_{m:} \right\|_2 \leq L \sqrt{p_0} \left\| W \right\|_F \sqrt{\log(n)} + L \left\| W \right\|_{2,\infty} \log^2(n).
\]
Since \(\bar{\phi}(x) = \sqrt{\log(n)} L \frac{\sqrt{p_0}}{\lambda_1^*} x + L \frac{\log^3 n}{\lambda_1^*}\), we have
\[
L \sqrt{p_0} \left\| W \right\|_F \sqrt{\log(n)} + L \left\| W \right\|_{2,\infty} \log^2(n) \lesssim \Delta^* \left\| W \right\|_{2,\infty} \frac{\left\| W \right\|_F}{\sqrt{n} \left\| W \right\|_{2,\infty}}.
\]
This finishes the proof. \(\square\)

After showing that Assumptions A1–A4 hold, we can prove the following theorem.

**Proposition 2.** Let \(t := (g(\alpha^p) p_0^* + (1 - p_0^*)) p_0\). Suppose \(p_0 \geq \frac{\log^3 n}{n}\) and \(\sqrt{np_0 \log(n)} L \kappa^2 \leq C t \lambda_1^*\) for some known constant \(C\). Then, with probability \(1 - O(n^{-2})\), the following holds
\[
\left\| \bar{U} \right\|_{2,\infty} \lesssim \left\| \bar{U}^* \right\|_{2,\infty} + \frac{\sqrt{np_0 \kappa L}}{\lambda_1^* t} \left\| M^* \right\|_{2,\infty} / \lambda_1^*
\]
\[
\left\| \bar{U} \text{sgn}(H) - \bar{U}^* \right\|_{2,\infty} \lesssim \frac{\sqrt{np_0 \log(n)}}{t \lambda_1^*} \left\| \bar{U}^* \right\|_{2,\infty} + \frac{\sqrt{np_0 \kappa L}}{\lambda_1^* t} \left\| M^* \right\|_{2,\infty} / \lambda_1^*.
\]

**Proof.** Let \(\gamma = (C_1 + 1) \bar{\gamma}, \phi(x) = C_2 \bar{\phi}(x)\) where \(C_1, C_2\) are constants defined in Lemmas 11 and 13 respectively.

One can verify that \(\gamma = (C_1 + 1) \frac{\sqrt{np_0 \log(n)} L + \log^2(n) \kappa}{\lambda_1^* t}\). In order to apply Lemma 6, we still need to show that \(32 \kappa \max(\gamma, \phi(\gamma)) \leq 1\). Because \(p_0 \geq \frac{\log^3 n}{n}\) and \(\sqrt{np_0 \log(n)} L \kappa^2 \leq C t \lambda_1^*\), one can verify that \(32 \kappa \max(\gamma, \phi(\gamma)) \leq 1\) by choosing a sufficient small \(C\). Based on Lemma 11, Lemma 13, we can apply Lemma 6 and obtain that, with probability \(1 - O(n^{-2})\),
\[
\left\| \bar{U} \right\|_{2,\infty} \lesssim (\kappa + \phi(1)) \left\| \bar{U}^* \right\|_{2,\infty} + \gamma \left\| t M^* \right\|_{2,\infty} / \Delta^*
\]
\[
\left\| \bar{U} \text{sgn}(H) - \bar{U}^* \right\|_{2,\infty} \lesssim (\kappa(\kappa + \phi(1)) (\gamma + \phi(\gamma)) + \phi(1)) \left\| \bar{U}^* \right\|_{2,\infty} + \gamma \left\| t M^* \right\|_{2,\infty} / \Delta^*.
\]
Using the fact \(\Delta^* = t \lambda_1^*/\kappa, \phi(1) \leq 1 \leq \kappa\), we have
\[
\left\| \bar{U} \right\|_{2,\infty} \lesssim \kappa \left\| \bar{U}^* \right\|_{2,\infty} + \gamma \left\| M^* \right\|_{2,\infty} / \lambda_1^*
\]
\[
\left\| \bar{U} \text{sgn}(H) - \bar{U}^* \right\|_{2,\infty} \lesssim (\kappa^2 (\gamma + \phi(\gamma)) + \phi(1)) \left\| \bar{U}^* \right\|_{2,\infty} + \gamma \left\| M^* \right\|_{2,\infty} / \lambda_1^*.
\]
Plug in the definition of \(\gamma\) and \(\phi\), we complete the proof. \(\square\)
A.2. Asymmetric Case

Let $X_\Omega$ associated with $M^*, p^*_\lambda, \alpha^*, p_0$ be the observation generated by the model described in the Section 2. Let $t = (p^*_\lambda g(\alpha^*) + (1 - p^*_\lambda))p_0$. Let $M^* = U^*\Sigma^*V^*T, M = UV^T$ be the singular decomposition of $M^*$ and $M$, where $\hat{M} = \arg \min_{\text{rank}(M) \leq r} \|X' - M'\|_F$ and $X'$ is obtained from $X_\Omega$ by setting unobserved entries to 0. We construct the following: $\hat{M}^* := \begin{pmatrix} 0_{n \times n} & M^* \end{pmatrix}$, $\hat{M} = \begin{pmatrix} 0_{m \times m} & 0_{m \times m} \end{pmatrix}$.

One can verify that the spectral decomposition of $\hat{M}^*$ is

$$\hat{M}^* = \frac{1}{\sqrt{2}} \begin{pmatrix} U^* & V^* & -V^* \end{pmatrix} \begin{pmatrix} \Sigma^* & 0 & 0 \\ 0 & \Sigma^* & 0 \\ 0 & 0 & \Sigma^* \end{pmatrix} \begin{pmatrix} U^* & V^* & -V^* \end{pmatrix}^T.$$

Note that the largest $r$ singular values, $\sigma^*_1 \geq \sigma^*_2 \geq \ldots \geq \sigma^*_r$, of $M^*$ are the same as the largest $r$ eigenvalues of $\hat{M}^*$. The $(r + 1)$-th eigenvalue of $\hat{M}^*$ is non-positive. Let $\hat{U}^* = \frac{1}{\sqrt{2}} \begin{pmatrix} U^* \\ V^* \\ -V^* \end{pmatrix}$ be the eigenvectors associated with the largest $r$ singular values of $\hat{M}^*$. Similarly, let $\hat{X} := \begin{pmatrix} 0_{n \times n} & X \end{pmatrix}$. Let $\hat{U} = \frac{1}{\sqrt{2}} \begin{pmatrix} U \\ V \end{pmatrix}$ be the eigenvectors associated with the largest $r$ singular values of $\hat{X}$.

We can apply Proposition 2 to the $\hat{M}^*$ and $\hat{X}$ constructed in this subsection. This gives us the following result.

**Proposition 3.** Let $H = \frac{1}{2}(U^T U + V^T V)$, $N = n + m, \mu = \max \left( N \|U^*\|_{2,\infty}, N \|V^*\|_{2,\infty} \right) / r, \kappa = \sigma^*_1 / \sigma^*_r, t = (p^*_\lambda g(\alpha^*) + (1 - p^*_\lambda))p_0$. Suppose $p_0 \geq \frac{\log m}{m}$ and $\sqrt{mp_0 \log(m)}L\kappa \leq Ct\sigma^*_1$ for some known small constant $C$. Then, with probability $1 - O((nm)^{-1})$, the following holds

$$\|\|U\|_{2,\infty} \vee \|V\|_{2,\infty}\| \leq \kappa \sqrt{\frac{\mu r}{m}} \quad (9)$$

$$\left(\|U_{\text{sgn}}(H) - U^*\|_{2,\infty} \vee \|V_{\text{sgn}}(H) - V^*\|_{2,\infty}\right) \leq \frac{\sqrt{p_0 \log(m)\kappa^3 L \sqrt{\mu r}}}{\tau \sigma^*_1} \quad (10)$$

$$\|M' - tM^*\|_{\max} \leq \kappa^r \mu r L \sqrt{\frac{p_0 \log(m)}{m}}. \quad (11)$$

**Proof.** Note that $\sqrt{3} \|\hat{U}^*\|_{2,\infty} = (\|U^*\|_{2,\infty} \vee \|V^*\|_{2,\infty}) \leq \sqrt{\mu r / N}$. Furthermore, we have

$$\|\hat{M}^*\|_{2,\infty} = \|M^*\|_{2,\infty} \leq \|U^*\|_{2,\infty} \|\Sigma^*V^*\|_2 \leq \|U^*\|_{2,\infty} \sigma^*_1 \leq \sqrt{\mu r / N} \sigma^*_1.$$

Apply Proposition 2 on $\hat{M}^*$ and $\hat{X}$ along with the bound on $\|\hat{U}^*\|_{2,\infty}, \|\hat{M}^*\|_{2,\infty}$, we can obtain, with probability $1 - O(N^{-2})$,

$$\|\hat{U}\|_{2,\infty} \leq \kappa \sqrt{\frac{\mu r}{N}} \quad (12)$$

$$\|\hat{U}_{\text{sgn}}(H) - \hat{U}^*\|_{2,\infty} \leq \sqrt{\frac{N p_0 \log(N)}{\tau \sigma^*_1}} \kappa^3 L \sqrt{\mu r / N} = \frac{\sqrt{p_0 \log(N)\kappa^3 L \sqrt{\mu r}}}{\tau \sigma^*_1}. \quad (13)$$

This completes the proof of Eq. (9) and Eq. (10). Next we proceed to the proof of Eq. (11).

Let $\hat{U} = U_{\text{sgn}}(H), \hat{V} = V_{\text{sgn}}(H), \hat{\Sigma} = \text{sgn}(H)^T \Sigma \text{sgn}(H)$. Note that $M'_{ij} = U_{i,} \Sigma V_{j,}^T = \hat{U}_{i,} \hat{\Sigma} \hat{V}_{j,}^T$ and $M^*_{ij} = U^*_{i,} \Sigma^* V^*_{j,}$. Then,

$$|M'_{ij} - tM^*_{ij}| = |\text{tr}(\hat{U}_{i,} \hat{\Sigma} \hat{V}_{j,}^T) - \text{tr}(U^*_{i,} \Sigma^* V^*_{j,})|$$

$$= |\text{tr}(\hat{\Sigma} \hat{V}_{j,}^T \hat{U}_{i,}) - \text{tr}(\Sigma^* V^*_{j,} U^*_{i,})|$$

$$= |\text{tr}(\hat{\Sigma} - \Sigma^*)(\hat{V}_{j,}^T \hat{U}_{i,}) + \text{tr}(\Sigma^* (\hat{V}_{j,}^T \hat{U}_{i,} - V^*_{j,}^T U^*_{i,}))|$$

$$\leq \|\hat{\Sigma} - \Sigma^*\|_2 \|\hat{V}_{j,}^T \hat{U}_{i,}\|_4 + t \|\Sigma^*\|_2 \|\hat{V}_{j,}^T \hat{U}_{i,} - V^*_{j,}^T U^*_{i,}\|_4 \quad (14)$$
where Eq. (14) is due to the triangle inequality and $|\text{tr}(AB)| \leq \|A\|_2 \|B\|_\infty$ by the Von Neumann’s trace inequality. We derive the bound on the term $\|\tilde{V}_j^T \tilde{U}_i, - V_j^* T U_i^* \|_\infty$. Let $\hat{\gamma} = \kappa \sqrt{\log(N)}(\sigma_l^* t)$. Note that

$$
\|\tilde{V}_j^T \tilde{U}_i, - V_j^* T U_i^* \|_\infty = \|(V_j^T - V_j^* T) \tilde{U}_i, + V_j^* T (\tilde{U}_i, - U_i^*)\|_\infty
\leq \|V_j^T - V_j^* T\|_2 \|\tilde{U}_i,\|_2 + \|V_j^* T (\tilde{U}_i, - U_i^*)\|_2
\leq \kappa^2 \sqrt{\log(N)} \hat{\gamma} \sqrt{\mu r / N} \left(\|\tilde{U}_i,\|_2 + \|V_j^* \|_2\right) \leq \kappa^3 \sqrt{\log(N)} \hat{\gamma} \mu r / N
$$

(16)

where Eq. (15) is due to $\|a b^T\|_\infty = \|a\|_2 \|b\|_2$ for any vector $a, b$. Eq. (16) is due to Eq. (13), and Eq. (17) is due to Eq. (12). We then bound $\|\tilde{V}_j^T \tilde{U}_i,\|_\infty$,

$$
\|\tilde{V}_j^T \tilde{U}_i,\|_\infty \leq \|V_j^* T U_i^*\|_\infty + \|\tilde{V}_j^T \tilde{U}_i, - V_j^* T U_i^*\|_\infty
\leq \|V_j^* T U_i^*\|_\infty + \kappa^3 \sqrt{\log(N)} \hat{\gamma} \mu r / N
\leq \|V^* \|_{2,\infty} \|U^*\|_{2,\infty} + \kappa^3 \sqrt{\log(N)} \hat{\gamma} \mu r / N
\leq \mu r / N + \kappa^3 \sqrt{\log(N)} \hat{\gamma} \mu r / N
\leq \kappa^2 \mu r / N
$$

(19)

where Eq. (18) is due to Eq. (17) and Eq. (19) is due to $\kappa \sqrt{\log(N)} \hat{\gamma} \lesssim 1$. Next we bound $\|\tilde{\Sigma} - \Sigma^*\|_2$. Note that

$$
\|\tilde{\Sigma} - \Sigma\|_2 = \|\text{sgn}(H)^T (\Sigma \text{sgn}(H) - \text{sgn}(H) \Sigma)\|_2
\leq \|\Sigma \text{sgn}(H) - \text{sgn}(H) \Sigma\|_2
= \|\Sigma H - H \Sigma\|_2 + \|(H - \text{sgn}(H)) \Sigma\|_2
\leq \|\Sigma H - H \Sigma\|_2 + 2 \|\Sigma\|_2 \|\text{sgn}(H) - H\|_2
\lesssim \kappa \mu r / N
$$

(20)

By Lemma 2 in (Abbe et al., 2017), we have

$$
\|\text{sgn}(H) - H\|_2 \lesssim (\|X - t\tilde{M}\|_2 / (t\sigma_l^*))^2 \lesssim \hat{\gamma}^2
\|\Sigma H - H \Sigma\|_2 \leq 2 \|X - t\tilde{M}\|_2 \lesssim \hat{\gamma} \sigma_r^*
$$

(21)

(22)

where $\|X - t\tilde{M}\|_2 \leq \hat{\gamma} t\sigma_r^*$ by Lemma 11. By Weyl’s inequality, we also have $\|\Sigma - \Sigma^*\|_2 \leq \|X - t\tilde{M}\|_2 \lesssim \hat{\gamma} \sigma_r^*$. Hence,

$$
\|\Sigma\|_2 \leq \|t\Sigma^*\|_2 + \|\Sigma - t\Sigma^*\|_2 \lesssim t\sigma_l^* + \hat{\gamma} \sigma_r^* \lesssim t\sigma_l^*.
$$

(23)

Plugging Eqs. (21) to (23) into Eq. (20), we have $\|\tilde{\Sigma} - \Sigma\|_2 \lesssim \hat{\gamma} \sigma^* + \hat{\gamma}^2 \sigma_l^* \lesssim \hat{\gamma} \sigma_1^*$. Therefore,

$$
\|\tilde{\Sigma} - \Sigma^*\|_2 \leq \|\tilde{\Sigma} - \Sigma\|_2 + \|\Sigma - \Sigma^*\|_2 \lesssim \hat{\gamma} \sigma_1^*.
$$

(24)

Plugging Eqs. (17), (19) and (24) into Eq. (14), we arrive at

$$
\left\|M^* - t\tilde{M}^*\right\|_{\max} \lesssim \kappa \sqrt{\log(N)} \hat{\gamma} \mu r / N \sigma_1^* \lesssim \kappa^3 \sqrt{\log(N)} \mu r / N \sigma_1^* \lesssim \kappa^4 \mu r / N \sqrt{\log(N)}.
$$

Next, we provide a lemma for the concentration bound of the sum over $\Omega$. 
Near-Optimal Anomaly Detection for Matrices with Sub-Exponential Noise

Lemma 14. Let $\Omega = \{(i,j) | O_{ij} = 1\} \subset [n] \times [m]$ where $O_{ij} \sim \text{Ber}(p_{ij})$ are i.i.d random variables. Let $\{T_{ij} | (i,j) \in [n] \times [m]\}$ be independent random variables with $E(T_{ij}) = p_{ij}$. Let

$$S = \sum_{(i,j) \in \Omega} T_{ij}.$$

Then, with probability $1 - 1/(nm)$,

$$\left| \sum_{(i,j) \in \Omega} T_{ij} - S \right| \leq C \left( \sqrt{S \log(nm)} + \log(nm) \right)$$

where $C$ is a constant. In particular, if $S \gtrsim \log(nm)$, then

$$\left| \sum_{(i,j) \in \Omega} T_{ij} - S \right| \leq C_1 S$$

where $C_1$ is a constant.

Proof. Let $Z_{ij} = T_{ij}O_{ij}$. Then $Z_{ij} \in [0, 1]$, $E(Z_{ij}) = p_{ij}p_{ij}$. By the Bernstein’s inequality (Bernstein, 1946), we have

$$P \left( \left| \sum_{ij} Z_{ij} - \sum_{ij} p_{ij} \right| > t \right) \leq 2e^{-\frac{t^2}{2E((Z_{ij} - p_{ij})^2)}} \leq 2e^{-\frac{t^2}{2}}$$

due to

$$E((Z_{ij} - p_{ij})^2) = E(Z_{ij}^2) \leq E(Z_{ij}) = p_{ij}p_{ij}.$$

Take $t = C_2 \left( \sqrt{S \log(nm)} + \log(nm) \right)$ where $C_2$ is a constant. Then we have

$$P \left( \left| \sum_{(i,j) \in \Omega} T_{ij} - \sum_{(i,j) \in [n] \times [m]} S \right| > t \right) \leq \frac{1}{nm}$$

for a proper $C_2$. \qed

Proof of Theorem 2. Next we proceed the proof of Theorem 2 based on Proposition 3. By the assumption in Section 2, $\log^{1/2}(\nu) \mu r Ln^2/(\| M^* \|_{\text{max}} \sqrt{m}) \lesssim \sqrt{p_0}$ and $1 - p_*^* \gtrsim 1$. Note that $\| M^* \|_{\text{max}} \lesssim \sigma_1^* \mu r \sqrt{m}$, this implies that $\sqrt{\log(m)} \sqrt{m} L n^2 \lesssim \sqrt{\nu} \sigma_1^*$ and $\sqrt{p_0} \gtrsim \sqrt{\log(m)} \sqrt{m}$, which is the condition required by Proposition 3. Also, by taking $T_{ij} = 1$ in Lemma 14 and noting that $p_{ij} \gtrsim \frac{\log(m)}{m}$, we have, with probability $1 - O(\frac{1}{nm})$,

$$\left| nmp_{ij} - \Omega \right| < C \sqrt{\log(nm)p_{ij}nm}$$

where $C$ is a constant. Then

$$\left| \frac{nm}{\Omega} - \frac{1}{p_{ij}} \right| = \frac{\left| nmp_{ij} - \Omega \right|}{\Omega p_{ij}} \leq \frac{C \sqrt{\log(nm)p_{ij}nm}}{\sqrt{p_{ij}nm}} \leq \frac{C'}{\sqrt{p_{ij}nm}}$$
where $C'$ is a constant. Finally, we can obtain
\[
\left\| \frac{M'_{nm}}{[\Omega]} - \frac{t}{p_O} M^* \right\|_{\max} = \left\| \frac{M'_{nm}}{[\Omega]} - M' \frac{1}{p_O} + M' \frac{1}{p_O} - \frac{t}{p_O} M^* \right\|_{\max} \\
\lesssim \left\| \frac{1}{p_O} (M' - tM^*) \right\|_{\max} + \left\| M' \right\|_{\max} \frac{\sqrt{\log(nm)/p_O nm}}{p_O} \\
\lesssim \frac{\kappa^4 \mu r L}{p_O} \sqrt{\frac{\log(m)p_O}{m}} + L \sqrt{\frac{\log(nm)}{p_O nm}} \\
\lesssim \kappa^4 \mu r L \sqrt{\frac{\log(m)}{p_O nm}}.
\] (25)

This completes the proof. \[ \square \]

B. Analysis of $\pi^{EW}$ and Proof of Theorem 1

B.1. Moment Matching Estimator

B.1.1. Proof of Lemma 1

Recall that
\[
g_t(\theta, M) = \frac{1}{nm} \sum_{(i,j) \in [n] \times [m]} (p_A \mathbb{P}_{\text{Anom}}(X_{ij} \leq t|\alpha, M_{ij}) + (1 - p_A) \mathbb{P}_{\text{Poisson}}(X_{ij} \leq t|M_{ij})).
\]

Let $\delta' = \kappa^4 \mu r L \sqrt{\frac{\log(m)}{p_O nm}}$ and
\[
h(\theta) = \sum_{t=0}^{T-1} \left( g_t(\theta, \hat{M}/e(\theta)) - \frac{|X_{ij} = t, (i,j) \in \Omega|}{|\Omega|} \right)^2.
\]

We have the following result.

**Lemma 15.** With probability $1 - O(\frac{1}{nm})$, for any $\theta \in \Theta$ and $t = 0, 1, \ldots, T$,
\[
|g_t(\theta, M^* e(\theta^*)/e(\theta)) - g_t(\theta, \hat{M}/e(\theta))| \lesssim (K + L)\delta'.
\]

**Proof.** Note that $\mathbb{P}_{\text{Anom}}(X_{ij} = t|\alpha, M)$ is $K$-lipschitz on $M$. One also can verify that $\mathbb{P}_{\text{Poisson}}(X_{ij} = t|M_{ij})$ is $L$-Lipschitz on $M$. Hence
\[
(p_A \mathbb{P}_{\text{Anom}}(X_{ij} \leq t|\alpha, M_{ij}) + (1 - p_A) \mathbb{P}_{\text{Poisson}}(X_{ij} \leq t|M_{ij}))
\]
is $(K + L)$-Lipschitz on $M_{ij}$. Let $C_1, C_2$ be two constants. By Theorem 2, with probability $1 - O((nm)^{-1})$, $|\hat{M}_{ij}/e(\theta^*) - M^*_{ij}| \leq C_1 \delta'$. This implies that
\[
\left| \frac{M^*_e(\theta^*)}{e(\theta)} - \frac{\hat{M}_{ij}}{e(\theta)} \right| \leq C_1 \delta' \leq C_2 \delta'
\]
where we use that $e(\theta) \geq (1 - p_A) \geq c$ for some constant $c$. This implies that
\[
\left| g_t \left( \theta, \frac{M^*_e(\theta^*)}{e(\theta)} \right) - g_t \left( \theta, \frac{\hat{M}_{ij}}{e(\theta)} \right) \right| \leq \frac{1}{nm} \sum_{ij} \left| \frac{M^*_e(\theta^*)}{e(\theta)} - \frac{\hat{M}_{ij}}{e(\theta)} \right| (K + L) \\
\lesssim (K + L)\delta'.
\]

\[ \square \]
Lemma 16. With probability $1 - O((nm)^{-1})$, $h(\theta^* ) \lesssim (K + L)^2 (\delta')^2$.

Proof. Set $C_1, C_2, C_3, C_4, C_5$ be proper constants.

Note that by Lemma 14, with probability $1 - O((nm)^{-1})$,
\[
||X_{ij} = t, (i,j) \in \Omega| - p_{O \text{num}}(\theta^*, M^*)|| \leq C_2 \sqrt{p_{O \text{num}}(\theta^*, M^*) \log(nm)} + C_2 \log(nm)
\]

Also, we can similarly obtain $||\Omega| - p_{O \text{num}}| \leq C_3 \sqrt{p_{O \text{num}} \log(nm)}$ by Lemma 14. Then, one can verify that
\[
\begin{align*}
|X_{ij} = t, (i,j) \in \Omega| - g_t(\theta^*, M^*) &= |X_{ij} = t, (i,j) \in \Omega| - |\Omega|g_t(\theta^*, M^*)| \\
&\leq \frac{1}{|\Omega|} \left( C_2 \sqrt{p_{O \text{num}}(\theta^*, M^*) \log(nm)} + C_3 \sqrt{p_{O \text{num}} \log(nm)}g_t(\theta^*, M^*) + C_2 \log(nm) \right) \\
&\leq C_4 \frac{\sqrt{p_{O \text{num}} \log(nm)}}{\sqrt{nmp_o}} \\
&\leq C_4 \frac{\log(nm)}{\sqrt{nmp_o}}.
\end{align*}
\]
Then, taking $\theta = \theta^*$ in Lemma 15, we have
\[
h(\theta^*) = \sum_{t=0}^T \left( g_t(\theta^*, \hat{M}/e(\theta^*)) - |X_{ij} = t, (i,j) \in \Omega|/|\Omega| \right)^2 \\
\leq \sum_{t=0}^T \left( |g_t(\theta^*, \hat{M}/e(\theta^*)) - g_t(\theta^*, M^*)| + C_5 \log(nm) / \sqrt{nmp_o} \right)^2 \\
\lesssim (K + L)^2 (\delta')^2 + \log(nm) \log(nm)/(nm) \\
\lesssim (K + L)^2 (\delta')^2
\]
due to the fact that $\delta' \gtrsim \sqrt{\frac{\log(nm)}{p_{O \text{num}}}}$ and $p_O \gtrsim \frac{\log^2(nm)}{m}$.
\[\square\]

Proof of Lemma 1. By Lemma 16, with probability $1 - O((nm)^{-1})$, $h(\hat{\theta}) \leq h(\theta^*) \lesssim (K + L)^2 (\delta')^2$. This implies, for each $t < T$, $|g_t(\hat{\theta}, \hat{M}/e(\hat{\theta})) - |X_{ij} = t, (i,j) \in \Omega|/|\Omega| | \lesssim (K + L)\delta'$. Combining with Lemma 16, we have, for each $t < T$,
\[
|g_t(\hat{\theta}, \hat{M}/e(\hat{\theta})) - g_t(\theta^*, \hat{M}/e(\theta^*))| \lesssim (K + L)\delta'.
\]
(26)

Note that
\[
|g_t(\theta^*, M^*) - g_t(\hat{\theta}, M^*)/e(\theta^*)/e(\hat{\theta})| \\
\leq |g_t(\theta^*, M^*) - g_t(\hat{\theta}, \hat{M}/e(\hat{\theta}))| + |g_t(\hat{\theta}, \hat{M}/e(\hat{\theta})) - g_t(\hat{\theta}, M^*/e(\theta^*)/e(\hat{\theta}))|.
\]
By Lemma 15, $|g_t(\hat{\theta}, \hat{M}/e(\hat{\theta})) - g_t(\theta^*, M^*)/e(\theta^*)/e(\hat{\theta})| \lesssim (K + L)\delta'$. Also we have
\[
|g_t(\theta^*, M^*) - g_t(\hat{\theta}, \hat{M}/e(\hat{\theta}))| \leq |g_t(\theta^*, M^*) - g_t(\theta^*, \hat{M}/e(\theta^*))| + |g_t(\hat{\theta}, \hat{M}/e(\hat{\theta})) - g_t(\theta^*, \hat{M}/e(\theta^*))|.
\]
By Lemma 15 again, we have $|g_t(\theta^*, M^*) - g_t(\theta^*, \hat{M}/e(\theta^*))| \lesssim (K + L)\delta'$. By Eq. (26), we have $|g_t(\hat{\theta}, \hat{M}/e(\hat{\theta})) - g_t(\theta^*, \hat{M}/e(\theta^*))| \lesssim (K + L)\delta'$. In conclusion,
\[
|g_t(\theta^*, M^*) - g_t(\hat{\theta}, \hat{M}/e(\hat{\theta}))| \lesssim (K + L)\delta'.
\]
Therefore, $\| F(\hat{\theta}) - F(\theta^*) \| \lesssim (K + L)\delta'$ since $T$ is a constant. \[\square\]
B.1.2. Proof of Lemma 2

Lemma 17. Suppose $F$ satisfies the following condition:

- $F : \Theta \subset \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ is continuously differentiable and injective.
- $B_{2C_2\delta}(\theta^*) \subset \Theta$ where $B_r(\theta^*) = \{ \theta : \| \theta - \theta^* \| \leq r \}$.
- $\| J_F(\theta) - J_F(\theta^*) \|_{\text{max}} \leq C_1 \| \theta - \theta^* \|$ for $\theta \in B_{2C_2\delta}(\theta^*)$.
- $\| J_F(\theta^*)^{-1} \|_2 \leq C_2$.

Suppose $2\sqrt{d_1d_2}C_1(C_2)^2\delta < 1/2$. For any $\theta \in \Theta$,

$$\| F(\theta) - F(\theta^*) \| \leq \delta \implies \| \theta - \theta^* \| \leq 2C_2\delta.$$  \hfill (27)

Proof. Suppose $\| F(\theta) - F(\theta^*) \| \leq \delta$. We construct a sequence of $\theta_i$ such that $\lim_{i \to \infty} F(\theta_i) = F(\theta)$ while $\| \theta_i - \theta^* \|$ is well bounded for every $i$. Let $\theta_i - \theta^* = J^{-1}_F(\theta^*)(F(\theta) - F(\theta^*))$. Note that

$$\| \theta_i - \theta^* \| \leq \| J^{-1}_F(\theta^*)\|_2 \| F(\theta) - F(\theta^*) \| \leq C_2\delta.$$  

Furthermore, by multivariate Taylor theorem,

$$F(\theta_i) = F(\theta^*) + A(\theta_i - \theta^*)^T$$

where the $i$-th row $A_i = (\nabla F_i(x_i))^T$ such that $x_i = \theta^* + c(\theta_i - \theta^*)$ for some $c \in [0,1]$. Hence, $F(\theta_i) = F(\theta^*) + J_F(\theta^*)(\theta_i - \theta^*)^T + (A - J_F(\theta^*))(\theta_i - \theta^*)^T$. Note that $F(\theta^*) + J_F(\theta^*)(\theta_i - \theta^*)^T = F(\theta)$ by the definition of $\theta_i$. Therefore,

$$\begin{align*}
F(\theta_i) &= F(\theta) + (A - J_F(\theta^*))(\theta_i - \theta^*)^T \\
\implies \| F(\theta_i) - F(\theta) \| &\leq \| A - J_F(\theta^*) \|_2 \| \theta_i - \theta^* \| \\
\implies \| F(\theta_i) - F(\theta) \| &\leq \| A - J_F(\theta^*) \|_{\text{max}} \sqrt{d_1d_2} \| \theta_i - \theta^* \| \\
\implies \| F(\theta_i) - F(\theta) \| &\leq C_1 \sqrt{d_1d_2} \| \theta_i - \theta^* \|^2 \\
\implies \| F(\theta_i) - F(\theta) \| &\leq C_1 \sqrt{d_1d_2} C_2^2 \delta^2.
\end{align*}$$

We can use the similar idea to the successive construction. In particular, let $t = 2\sqrt{d_1d_2}C_1(C_2)^2\delta < 1/2, a = 2C_1C_2\sqrt{d_1d_2}, \theta_0 = \theta^*$. Suppose

$$\| \theta_k - \theta_{k-1} \| \leq \frac{1}{a} t^k, \| \theta_k - \theta^* \| \leq \frac{1}{a} (2t - t^k), \| F(\theta_k) - F(\theta) \| \leq \frac{1}{aC_2} t^{k+1}. $$

It is easy to verify that the above conditions are satisfied for $k = 1$. Then let $\theta_{k+1} - \theta_k = J^{-1}_F(\theta^*)(F(\theta) - F(\theta_k))$ for $k > 1$.

Then, we have $\| \theta_{k+1} - \theta_k \| \leq C_2 \frac{t^{k+1}}{aC_2} \leq \frac{t^{k+1}}{a}$. Also, $\| \theta_{k+1} - \theta^* \| \leq \| \theta_{k+1} - \theta_k \| + \| \theta_k - \theta^* \| \leq \frac{2t - t^k + t^{k+1}}{a} \leq \frac{2t - t^k}{a}$. Furthermore,

$$\begin{align*}
F(\theta_{k+1}) &= F(\theta_k) + J_F(\theta^*)(\theta_{k+1} - \theta_k)^T + (A - J_F(\theta^*))(\theta_{k+1} - \theta_k)^T \\
\implies F(\theta_{k+1}) &= F(\theta) + (A - J_F(\theta^*))(\theta_{k+1} - \theta_k)^T \\
\implies \| F(\theta_{k+1}) - F(\theta) \| &\leq \| A - J_F(\theta^*) \|_2 \| \theta_{k+1} - \theta_k \| \\
\implies \| F(\theta_{k+1}) - F(\theta) \| &\leq \| A - J_F(\theta^*) \|_{\text{max}} \sqrt{d_1d_2} \| \theta_{k+1} - \theta_k \| \\
\implies \| F(\theta_{k+1}) - F(\theta) \| &\leq C_1 \sqrt{d_1d_2} \| \theta_{k+1} - \theta_k \| \\
\implies \| F(\theta_{k+1}) - F(\theta) \| &\leq C_1 \sqrt{d_1d_2} \frac{2(t) \alpha t^{k+1}}{a} \\
\implies \| F(\theta_{k+1}) - F(\theta) \| &\leq C_1 C_2 \sqrt{d_1d_2} \frac{t^{k+2}}{aC_2} \leq \frac{t^{k+2}}{aC_2}.
\end{align*}$$
Note that \( \| \theta_k - \theta^* \| \leq \frac{2\alpha}{a} \). Therefore \( \theta_k \in \Theta \) is well-defined. Furthermore, we can conclude for any \( \epsilon > 0 \), there exists \( N \), if \( k_1, k_2 > N \), \( \| \theta_{k_1} - \theta_{k_2} \| \leq \epsilon \). Therefore, the sequence converges. Suppose \( \lim_{k} \theta_k = \theta' \). Note that \( \| \theta' - \theta^* \| \leq \frac{2\alpha}{a} \) due to \( \| \theta_k - \theta^* \| \leq \frac{2\alpha}{a} \). Also note that \( \| J_F(\theta) \| \) is bounded for \( \| \theta - \theta^* \| \leq \frac{2\alpha}{a} \). This implies that \( \lim_{k} F(\theta_k) = F(\theta') \). On the other hand, due to the convergence of \( F(\theta_k) \), \( \lim_{k} F(\theta_k) = F(\theta) \). By injectivity, \( \theta = \theta' \) and \( \| \theta - \theta^* \| \leq 2t/a \). This completes the proof. 

**Proof of Lemma 2.** We proceed the proof of Lemma 2. By the regularity conditions (RC) stated in Section 4.2 and Lemma 17, we have \( \| F(\theta) - F(\theta^*) \| \lesssim (K + L)\delta' \implies \| \theta - \theta^* \| \lesssim (K + L)\delta' \). By Lemma 1, we complete the proof. 

**B.2. Confidence Interval Estimator and Proof of Lemma 3**

Let \( \delta' = (K + L)\kappa^4 \eta L \sqrt{\frac{\log(m)}{m \alpha^2}} \). Let

\[
\hat{x}_{ij} := [\hat{p}_A p_{\text{Anom}}(X_{ij} | \hat{\alpha}, \hat{M}_{ij} / e(\hat{\theta}))]
\]

\[
\hat{y}_{ij} := [(1 - \hat{p}_A) p_{\text{Poisson}}(X_{ij} | \hat{M}_{ij} / e(\hat{\theta}))].
\]

Let \( x_{ij} = p_{\text{Anom}} A_{\text{Anom}} \left( X_{ij} | \alpha*, M_{ij}^* \right) \), \( y_{ij} = (1 - p_{\text{Anom}}) p_{\text{Poisson}} \left( X_{ij} | M_{ij}^* \right) \). We have the following result.

**Lemma 18.** With probability \( 1 - O(1/(nm)) \), \( \max(|\hat{x}_{ij} - x_{ij}|, |\hat{y}_{ij} - y_{ij}|) \leq C(L + K)^2 L \delta' \) for any \( (i, j) \in \Omega \).

**Proof.** By Lemma 2, with probability \( 1 - O(1/(nm)) \), we have \( \| \hat{\theta} - \theta^* \| \lesssim \delta' \).

Note that \( g(\theta) \) is \( K \)-Lipschitz in \( \theta \) and \( e(\theta) = p_{\text{Anom}} g(\theta) + (1 - p_{\text{Anom}}) \). Hence

\[
|e(\hat{\theta}) - e(\theta^*)| \leq |\hat{p}_A - p_{\text{Anom}}| (1 - g(\hat{\theta})) + p_{\text{Anom}}|g(\hat{\theta}) - g(\theta^*)| \lesssim (K + 1)\delta'.
\]

Furthermore

\[
\frac{M_{ij}}{e(\theta)} - M^* \leq \frac{1}{e(\theta)} |\hat{M} - M^* e(\hat{\theta})| \leq \frac{1}{e(\theta)} \left( |\hat{M} - M^* e(\theta^*)| + M^* |e(\theta^*) - e(\hat{\theta})| \right) \lesssim \frac{\delta'}{K + L} + L(K + 1)\delta' \lesssim L(K + 1)\delta'.
\]

Note that \( p_{\text{Anom}}(\alpha, M) \) is \( K \)-Lipschitz in \( \alpha \) and \( M \). The implies that

\[
|\hat{x}_{ij} - x_{ij}| \leq |\hat{p}_A p_{\text{Anom}} \left( X_{ij} | \hat{\alpha}, \hat{M}_{ij} \right) - p_{\text{Anom}} X_{ij} | \alpha^*, M_{ij}^* \right) | \leq |\hat{p}_A - p_{\text{Anom}}| X_{ij} | \alpha^*, M_{ij}^* \right) + |p_{\text{Anom}} X_{ij} | \alpha^*, M_{ij}^* \right) - p_{\text{Anom}} \left( X_{ij} | \hat{\alpha}, \hat{M}_{ij} \right) | \hat{p}_A \lesssim \delta' + KL(K + 1)\delta' \lesssim KL(K + 1)\delta'.
\]

Similarly, one can obtain \( |\hat{y}_{ij} - y_{ij}| \lesssim L^2(K + 1)\delta' \). In conclusion,

\[
\max(|\hat{x}_{ij} - x_{ij}|, |\hat{y}_{ij} - y_{ij}|) \lesssim (L + K)^2 L.
\]
Lemma 19. Suppose $|\hat{x} - x| \leq \delta, |\hat{y} - y| \leq \delta$ where $x, y, \hat{x}, \hat{y} \in [0, 1], x + y > 0$. Let $\hat{s} = \frac{\hat{x}}{x + y}$ if $\hat{x} + \hat{y} > 0$ otherwise $\hat{s} = 0$. Then,

$$\left| \hat{s} - \frac{x}{x + y} \right| \leq \min \left( \frac{\delta}{x + y}, \frac{\delta}{\hat{x} + \hat{y}}, 1 \right).$$

(28)

Proof. Note that $\left| \hat{s} - \frac{x}{x+y} \right| \leq 1$ is trivial since $\hat{s} \in [0, 1]$ and $\frac{x}{x+y} \in [0, 1]$.

When $\hat{x} = \hat{y} = 0$, $\frac{x}{x+y} \leq \frac{\delta}{x+y}$ due to $x \leq \delta$.

When $\hat{x} + \hat{y} > 0$,

$$\left| \hat{s} - \frac{x}{x+y} \right| = \frac{\hat{x}(x+y) - x(\hat{x} + \hat{y})}{(\hat{x} + \hat{y})(x+y)} = \left| \frac{\hat{x}y - xy\hat{y}}{(\hat{x} + \hat{y})(x+y)} \right| = \left| \frac{\hat{x}(\hat{y} - (y - y)) - (\hat{x} - (\hat{x} - x))\hat{y}}{(\hat{x} + \hat{y})(x+y)} \right| = \left| \frac{-\hat{x}(\hat{y} - y) + (\hat{x} - x)\hat{y}}{(\hat{x} + \hat{y})(x+y)} \right| \leq \frac{\hat{x}}{\hat{x} + \hat{y}} \frac{|\hat{y} - y|}{x + y} + \frac{\hat{y}}{\hat{x} + \hat{y}} \frac{|\hat{x} - x|}{x + y} \leq \frac{\hat{x}}{\hat{x} + \hat{y}} \frac{\delta}{x + y} + \frac{\hat{y}}{\hat{x} + \hat{y}} \frac{\delta}{x + y} = \frac{\delta}{x+y}.$$

By symmetry, $\left| \hat{s} - \frac{x}{x+y} \right| \leq \frac{\delta}{x+y}$, which completes the proof.

Proof of Lemma 3. Next, we proceed the proof of Lemma 3. For notation simplification, write $\hat{x}_{ij}, \hat{y}_{ij}, x_{ij}, y_{ij}$ as $\hat{x}, \hat{y}, x, y$.

Let $\delta = (K + L)^3 \kappa^4 \mu r L^2 \sqrt{\log(\mathrm{exp})} \frac{1}{\rho \sigma m}$ and $C$ be the constant denoted in Lemma 18. Then by Lemma 18, with probability $1 - O(1/(nm))$,

$$|x - \hat{x}| \leq C\delta, |y - \hat{y}| \leq C\delta.$$

By Lemma 19, we have

$$\frac{\hat{x} - C\delta}{\hat{x} + \hat{y}} \leq \frac{x}{x+y}.$$

Therefore $f_{ij}^1 \leq f_{ij}^*$. Next we show that $f_{ij}^* \leq f_{ij}^1 + 4C\delta$.

If $4C\delta \geq (x + y)$, then $f_{ij}^1 + 4C\delta \geq 1 \geq \frac{x}{x+y}$. On the other hand, suppose $4C\delta < (x + y)$. Note that

$$x + y > 4C\delta \implies 4(x + y - 2C\delta) > 2(x + y) \implies \frac{4C\delta}{x + y} > \frac{2C\delta}{x + y - 2C\delta} \implies \frac{4C\delta}{x + y} > \frac{2C\delta}{\hat{x} + \hat{y}}.$$
Then,
\[
\frac{x - C\delta}{x + y} + \frac{4C\delta}{x + y} \geq \left( \frac{x}{x + y} - \frac{2C\delta}{x + y} \right) + \frac{4C\delta}{x + y} \geq \frac{x}{x + y}.
\]
This implies that \( f_{ij}^L + \frac{4C\delta}{x + y} \geq f_{ij}^* \). Similar result can be shown for \( f_{ij}^R \). This completes the proof.

### B.3. Analysis of the optimization problem \( \mathcal{P}^{\text{EW}} \)

Note that \( \mathcal{P}^{\text{EW}} \) is obtained from \( \mathcal{P}^* \) by replacing \( f_{ij}^* \) with the confidence interval estimators \( f_{ij}^L \) and \( f_{ij}^R \). Intuitively, we could expect that \( \mathcal{P}^{\text{EW}} \approx \mathcal{P}^* \), and therefore the algorithm \( \pi^{\text{EW}} \) should achieve the desired performance. We first have the following lemma to show that \( \text{FPR}_{\pi^{\text{EW}}(\gamma)}(X) \leq \gamma \) since \( f_{ij}^L \leq f_{ij}^* \leq f_{ij}^R \) and so \( \{f_{ij}^{\text{EW}}\} \) is a feasible solution of \( \mathcal{P}^* \).

**Lemma 20.** With probability \( 1 - O(1/(nm)) \), for any \( 0 < \gamma \leq 1 \)
\[
\text{FPR}_{\pi^{\text{EW}}(\gamma)}(X_{\Omega}) \leq \gamma.
\]

**Proof.** This is because
\[
\sum_{(i,j) \in \Omega} f_{ij}^L \leq \sum_{(i,j) \in \Omega} f_{ij}^{\text{EW}} \leq \gamma \sum_{(i,j) \in \Omega} f_{ij}^L \leq \gamma \sum_{(i,j) \in \Omega} f_{ij}^*,
\]
due to that \( f_{ij}^L \leq f_{ij}^* \leq f_{ij}^R \) and the constraint of \( f_{ij}^{\text{EW}} \).

To show the desired performance guarantee for \( \text{TPR}_{\pi^{\text{EW}}}(X) \), we provide the following Lemma that characterizes how \( f_{ij}^L \) and \( f_{ij}^R \) are close to \( f_{ij}^* \) in an accumulated manner (the proof is shown momentarily):

**Lemma 21.** Let \( \delta = (K + L)^3 \kappa^4 \mu r L^2 \sqrt{\log(m)/p_0 m} \) With probability \( 1 - O(1/(nm)) \),
\[
\sum_{(i,j) \in \Omega} (|f_{ij}^L - f_{ij}^*| + |f_{ij}^R - f_{ij}^*|) \leq \frac{C L \log(m)}{\gamma p_A}.
\]

Next we proceed to the analysis of \( \mathcal{P}^{\text{EW}} \). For a fixed \( \eta \), let \( \{t_{ij}^\prime\} \) be the optimal solution of \( \pi^*(\gamma') \) for some \( \gamma' \) such that \( \sum_{(i,j) \in \Omega} t_{ij}^\prime = \eta < 1 \). The key idea is to find some \( \eta \) such that \( \{t_{ij}^\prime\} \) is a feasible solution of \( \mathcal{P}^{\text{EW}} \), while maintaining good TPR performance compared to \( \pi^*(\gamma) \). Indeed, a sufficiently large \( \eta \) can be achieved by Lemma 21. In particular, we have the (proof is shown momentarily):

**Lemma 22.** Let \( \delta = (K + L)^3 \kappa^4 \mu r L^2 \sqrt{\log(m)/p_0 m} \), \( \eta = 1 - CL\delta \log(m)/\gamma \). Then \( \{t_{ij}^\prime\} \) is a feasible solution of \( \mathcal{P}^{\text{EW}} \).

Furthermore, \( \min\left(1, \frac{\sum_{(i,j) \in \Omega} t_{ij}^\prime - \sum_{(i,j) \in \Omega} t_{ij}^\prime}{\sum_{(i,j) \in \Omega} (1 - t_{ij}^\prime)} \right) \leq C_1 \frac{L \delta \log(m)}{\gamma p_A} \) for a constant \( C_1 \).

### B.3.1. Proof of Lemma 21

Next, we prove Lemma 21, i.e., show that the accumulated error induced by the approximation of \( f_{ij}^* \) by \( f_{ij}^L \) and \( f_{ij}^R \) has the desired bound.

**Proof of Lemma 21.** Let \( x_{ij} := p_{A,\text{anom}}(X_{ij}|\alpha^*, M_{ij}^*) \), \( y_{ij} := (1 - p_{A,\text{anom}})\mathbb{P}_{\text{Poisson}}(X_{ij}|M_{ij}^*) \). By Lemma 3,
\[
\max(|f_{ij}^L - f_{ij}^*|, |f_{ij}^R - f_{ij}^*|) \leq \epsilon_{ij}
\]
where \( \epsilon_{ij} := \min\left(\frac{4C\delta}{x_{ij} + y_{ij}}, 1\right) \) for some constant \( C \) and \( \delta = (K + L)^3 \kappa^4 \mu r L^2 \sqrt{\log(m)/p_0 m} \).
Note that when $X_{ij} = t$,

$$x_{ij} + y_{ij} = \mathbb{P}(X_{ij} = t).$$

Note that $\|X_{ij}\|_{\psi_1} \leq L$ is a sub-exponential random variable by Lemmas 8 and 9. Then, we have

$$\mathbb{P}(X_{ij} > t) \leq \exp^{-t/C'L} \implies \mathbb{P}(X'_{ij} > C'L \log(1/\delta)) \leq \delta$$

where $C'$ is a proper constant. Let $z_{ij} = \min\left(\frac{\delta}{\mathbb{P}(X_{ij} = t)}, 1\right)$. Then,

$$E(z_{ij}) = \sum_{t=0}^{\infty} \min(1, \delta/\mathbb{P}(X_{ij} = t)) \mathbb{P}(X_{ij} = t) \leq C'L \log(1/\delta) \delta + \sum_{t=C'L \log(1/\delta) + 1}^{\infty} \mathbb{P}(X_{ij} = t) \leq C'L \log(1/\delta) \delta + \delta.$$

Note that $z_{ij} \in [0, 1]$ are independent random variables. Then, by Lemma 14, with probability $1 - O(\frac{1}{nm})$,

$$\sum_{(i,j) \in \Omega} z_{ij} \leq L \log(1/\delta) \delta p_{nm} + \sqrt{p_{nm} \log(nm)} \leq L \log(m) \delta p_{nm}$$

given that $\delta \gtrsim \sqrt{\frac{\log(nm)}{p_{nm}}}$.

Therefore,

$$\sum_{(i,j) \in \Omega} \max(|f^l_{ij} - f^*_ij|, |f^r_{ij} - f^*_ij|) \leq \sum_{(i,j) \in \Omega} \epsilon_{ij} \leq \sum_{(i,j) \in \Omega} z_{ij} \lesssim L \log(m) \delta p_{nm}.$$

\[\square\]

**B.3.2. Proof of Lemma 22**

Consider a concentration bound

**Lemma 23.** Let $C_1, C_2, C_3$ be constants. With probability $1 - O(\frac{1}{nm})$,

$$\sum_{(i,j) \in \Omega} f^*_{ij} \geq C_1 nm p_{nm}$$

$$|\Omega| \leq C_2 nm p_{nm}.$$

Furthermore, if $p_{nm} \gtrsim \log(nm)$,

$$\sum_{(i,j) \in \Omega} 1 - f^*_ij \geq C_3 p_{nm} \log(nm).$$

**Proof.** Let $Z_{ij} = \mathbb{P}(B_{ij} = 1|X_{ij})$. Then $\sum_{(i,j) \in \Omega} Z_{ij} = \sum_{(i,j) \in \Omega} Z_{ij}$. Note that $E(Z_{ij}) = p^*_A$ and $Z_{ij} \in [0, 1]$ are independent. Hence, by Lemma 14, with probability $1 - O(\frac{1}{nm})$, $\sum_{(i,j) \in \Omega} 1 - f^*_ij \geq C p^*_A p_{nm}$ where $C$ is a constant given that $p^*_A p_{nm} \gtrsim \log(nm)$ Similar results for $\sum_{(i,j) \in \Omega} f^*_ij$ (with $1 - p^*_A \geq c$ for some constant $c$) and $|\Omega|$ can also be obtained. \[\square\]
Proof of Lemma 22. Let \( \{t'_{ij}, (i,j) \in \Omega \} \) be the optimal solution of the algorithm \( \pi^*(\gamma') \). Let \( \{t^*_{ij}, (i,j) \in \Omega \} \) be the optimal solution of \( \pi^*(\gamma) \). Suppose
\[
\sum_{(i,j) \in \Omega} t'_{ij} = \eta < 1.
\]

Order \( f^*_{ij} \) by \( f^*_{a_1b_1} \leq f^*_{a_2b_2} \leq \ldots \leq f^*_{a_{|\Omega|}b_{|\Omega|}} \). One can easily verify that \( t'_{a_1b_1} \leq t^*_{a_1b_1}, t'_{a_2b_2} \leq t^*_{a_2b_2}, \ldots, t'_{a_{|\Omega|}b_{|\Omega|}} \leq t^*_{a_{|\Omega|}b_{|\Omega|}} \). Furthermore, for any \( k \) and \( l \) such that \( t'_{a_kb_k} > 0 \) and \( t^*_{a_kb_k} > 0 \), we have \( f^*_{a_kb_k} \leq f^*_{a_kb_k} \). Let
\[
A = \sum_{ij} t'_{ij}, B = \sum_{ij} t^*_{ij} - t'_{ij}, C = \sum_{ij} t'_{ij}^* f^*_{ij}, D = \sum_{ij} (t^*_{ij} - t'_{ij}) f^*_{ij}.
\]
Then the following weighted average inequality holds: \( \frac{C}{\eta} \leq \frac{B}{1-\eta} \). This implies that \( \frac{C}{\eta} \leq \frac{C+B}{\eta+1} \), i.e.,
\[
\frac{1}{\sum_{(i,j) \in \Omega} t'_{ij}} \sum_{(i,j) \in \Omega} t'_{ij} f^*_{ij} \leq \frac{1}{\sum_{(i,j) \in \Omega} t^*_{ij}} \sum_{(i,j) \in \Omega} t^*_{ij} f^*_{ij}.
\]

This implies that \( \sum_{(i,j) \in \Omega} t'_{ij} f^*_{ij} \leq \eta \sum_{(i,j) \in \Omega} t^*_{ij} f^*_{ij} \). Then, we have,
\[
\sum_{(i,j) \in \Omega} t'_{ij} f^*_{ij} \leq \sum_{(i,j) \in \Omega} t'_{ij} (f^*_{ij} + |f^R_{ij} - f^*_{ij}|) \leq \left( \eta \sum_{(i,j) \in \Omega} t'_{ij} f^*_{ij} \right) + \sum_{(i,j) \in \Omega} |f^R_{ij} - f^*_{ij}| \leq \gamma \sum_{(i,j) \in \Omega} f^*_{ij} + \sum_{(i,j) \in \Omega} |f^R_{ij} - f^*_{ij}| - \gamma(1-\eta) \sum_{(i,j) \in \Omega} f^*_{ij}.
\]

Note that
\[
\gamma \sum_{(i,j) \in \Omega} f^*_ij \leq \gamma \sum_{(i,j) \in \Omega} f^*_ij + \sum_{(i,j) \in \Omega} |f^*_ij - f^*_ij|.
\]

Therefore, we have
\[
\sum_{(i,j) \in \Omega} t'_{ij} f^*_ij \leq \gamma \sum_{(i,j) \in \Omega} f^*_ij + \left( \sum_{(i,j) \in \Omega} (|f^*_ij - f^*_ij| + |f^*_ij - f^R_{ij}|) \right) - \gamma(1-\eta) \sum_{(i,j) \in \Omega} f^*_ij.
\]

By Lemma 21, we have \( \left( \sum_{(i,j) \in \Omega} (|f^*_ij - f^*_ij| + |f^*_ij - f^R_{ij}|) \right) \leq C_1 L \log(m) \delta \rho_0 nm \). By Lemma 23, we have \( \gamma(1-\eta) \sum_{(i,j) \in \Omega} f^*_ij \geq C_2 (1-\eta) \rho_0 nm \). Take \( \eta = 1 - \frac{C_1}{C_2} L \log(m) \delta \). We then have \( \{t'_{ij}\} \) is a feasible solution of \( \mathcal{P}_{\text{EW}} \):
\[
\sum_{(i,j) \in \Omega} t'_{ij} f^R_{ij} \leq \gamma \sum_{(i,j) \in \Omega} f^R_{ij}.
\]

Furthermore, for any \( 0 < \gamma \leq 1 \), we can get
\[
\sum_{(i,j) \in \Omega} t'_{ij} - t^*_{ij} \leq |\Omega| \leq nmp_\Omega.
\]

By Lemma 23, \( \sum_{(i,j) \in \Omega} f^*_{ij} \leq |\Omega| \leq nmp_\Omega \). Suppose \( p_A^* \rho_0 nm \geq \log(nm) \), then by Lemma 23, \( \sum_{(i,j) \in \Omega} (1 - f^*_{ij}) \geq nmp_\Omega p_A^* \). This leads to
\[
\sum_{(i,j) \in \Omega} t'_{ij} - t^*_{ij} \leq \frac{(1-\eta) \rho_0 nm}{p_A^* \rho_0 nm} \leq \frac{L \log(m) \delta}{\gamma p_A^*}.
\]

(30)
Note that $\delta \gtrsim \frac{1}{\sqrt{np_{\theta}m}}$. Suppose $p_{\theta}A_{m} \gtrsim \log(nm)$, then

$$\frac{L \log(n)\delta}{\gamma p_{\theta}} \gtrsim \frac{1}{p_{\theta} \sqrt{p_{\theta}m}} \gtrsim \frac{nm \sqrt{p_{\theta}}}{\log(nm) \sqrt{m}} \gtrsim 1.$$ 

This completes the proof. \hfill \Box

B.4. Proof of Theorem 1

Proof of Theorem 1. Finally, we proceed the proof of Theorem 1. Note that

$$\text{TPR}_{\pi^{*}(\gamma)}(X_{\Omega}) - \text{TPR}_{\pi^{\text{EW}}(\gamma)}(X_{\Omega})$$

$$= \frac{\sum_{(i,j)\in \Omega} t_{ij}^* (1 - f_{ij}^*) - \sum_{(i,j)\in \Omega} t_{ij}^\text{EW} (1 - f_{ij}^*)}{\sum_{(i,j)\in \Omega} (1 - f_{ij}^*)}$$

$$\leq \frac{\sum_{(i,j)\in \Omega} (t_{ij}^* - t_{ij}^\text{EW}) + \sum_{(i,j)\in \Omega} t_{ij}^\text{EW} f_{ij}^* - \sum_{(i,j)\in \Omega} t_{ij}^* f_{ij}^*}{\sum_{(i,j)\in \Omega} (1 - f_{ij}^*)}.$$ 

Note that $\sum_{(i,j)\in \Omega} t_{ij}^* f_{ij}^* = \gamma \sum_{(i,j)\in \Omega} f_{ij}^*$ and $\sum_{(i,j)\in \Omega} t_{ij}^\text{EW} f_{ij}^* \leq \gamma \sum_{(i,j)\in \Omega} f_{ij}^*$ by Lemma 20. Furthermore, $\sum_{(i,j)\in \Omega} t_{ij}^\text{EW} \geq \sum_{(i,j)\in \Omega} t_{ij}^*$ since $\{ t_{ij}^* \}$ is a feasible solution of $P_{\text{EW}}$ and the objective function of $P_{\text{EW}}$ maximizes $\sum_{(i,j)\in \Omega} t_{ij}^\text{EW}$ given the constraint. Hence,

$$\text{TPR}_{\pi^{*}(\gamma)}(X_{\Omega}) - \text{TPR}_{\pi^{\text{EW}}(\gamma)}(X_{\Omega}) \leq \frac{\sum_{(i,j)\in \Omega} (t_{ij}^* - t_{ij}^\text{EW})}{\sum_{(i,j)\in \Omega} (1 - f_{ij}^*)} \leq \frac{\sum_{(i,j)\in \Omega} (t_{ij}^* - t_{ij}^\text{EW})}{\sum_{(i,j)\in \Omega} (1 - f_{ij}^*)}.$$ 

Also, note that $\text{TPR}_{\pi^{*}(\gamma)}(X_{\Omega}) - \text{TPR}_{\pi^{\text{EW}}(\gamma)}(X_{\Omega}) \leq 1$ since $\text{TPR} \leq 1$ by definition. By Lemma 22,

$$\text{TPR}_{\pi^{*}(\gamma)}(X_{\Omega}) - \text{TPR}_{\pi^{\text{EW}}(\gamma)}(X_{\Omega}) \lesssim \frac{L \log(m)\delta}{\gamma p_{\theta}},$$

which completes the proof. \hfill \Box

C. Minimax Lower Bound

C.1. Proof of Proposition 1

Consider the model $X \sim H(p_{\theta}^*, M^*)$. Recall that the construction of $M_n = \{M^b \in \mathbb{R}^{n \times n}, b \in \{0, 1\}^{n/2}\}$ is: for the $i$-th and $(i + 1)$-th rows, set $M_{ij} = 1$ and $M_{i+1,j} = 1 - \frac{e^*}{\sqrt{n}}$ if $b_{ij} = 0$; otherwise set $M_{ij} = 1 - \frac{e^*}{\sqrt{n}}$ and $M_{i+1,j} = 1$. Here $e^* < \frac{1}{2}$ is some sufficient small constant.

For a constant $C_0$, let $\Pi_c$ denote the set of all policies such that

$$\mathbb{P}_{X \sim H(p_{\theta}^*, M)}(\text{FPR}_{M}(X) \leq \gamma) \geq 1 - C_0/n^2 \quad \text{for all } M \in M_n. \quad (31)$$

Set $\gamma = \frac{1}{2e}, p_{\theta}^* = \frac{1}{2}$. We write $\sum_{(i,j)\in [n] \times [n]} a_{ij}$ as $\sum_{ij}$ if there is no ambiguity. Note that anomaly can only occur when $X_{ij} = 0$. One can verify that

$$f_{ij}^* = \frac{1/2e^{-M_{ij}}}{1/2 + 1/2e^{-M_{ij}}} \mathbb{1}\{X_{ij} = 0\} + \mathbb{1}\{X_{ij} > 0\}$$

$$= \frac{e^{-M_{ij}}}{1 + e^{-M_{ij}}} \mathbb{1}\{X_{ij} = 0\} + \mathbb{1}\{X_{ij} > 0\}.$$ 

Since the anomaly only occurs when $X_{ij} = 0$, a “rational” algorithm should not claim anomalies for those entries with $X_{ij} > 0$. We have the following result:
Lemma 24. For any \( \pi' \in \Pi'_\gamma \), there exists \( \pi \) such that for any \( X \),

\[
\text{FPR}_\pi(X) \leq \text{FPR}_{\pi'}(X), \text{TPR}_\pi(X) = \text{TPR}_{\pi'}(X).
\]

Furthermore, \( \mathbb{P} (A_{ij}^\pi = 1, X_{ij} > 0) = 0 \).

**Proof.** For any algorithm \( \pi' \), we can construct \( \pi \) as the following: let \( A_{ij}^\pi = A_{ij}^\pi' \) if \( X_{ij} = 0 \); otherwise \( A_{ij}^\pi = 0 \). Then it is easy to see that

\[
\sum_{ij} \mathbb{P} (A_{ij}^\pi = 1|X) f_{ij}^\pi \leq \sum_{ij} \mathbb{P} (A_{ij}^\pi' = 1|X) f_{ij}^\pi.
\]

This implies \( \text{FPR}_\pi(X) \leq \text{FPR}_{\pi'}(X) \). Furthermore,

\[
\sum_{ij} \mathbb{P} (A_{ij}^\pi = 1|X) (1 - f_{ij}^\pi) = \sum_{ij} \mathbb{P} (A_{ij}^\pi' = 1|X) (1 - f_{ij}^\pi).
\]

This implies \( \text{TPR}_\pi(X) = \text{TPR}_{\pi'}(X) \). \( \square \)

Hence, it is sufficient to only consider \( \pi \) that does not claim anomalies for entries with \( X_{ij} > 0 \). Let \( \Pi_\gamma = \{ \pi \in \Pi'_\gamma \mid \mathbb{P} (A_{ij}^\pi = 1, X_{ij} > 0) = 0 \} \). Note that the FPR constraint is a high probability statement, hence it is possible that different \( M \in \mathcal{M}_n \) satisfies the constraint on different sets of \( X \) and makes the problem hard to analyze. To address this issue, we consider the “expectation” of the FPR constraint and have the following lemma.

Lemma 25. For any \( \pi \in \Pi_\gamma \) and any \( M \in \mathcal{M} \),

\[
\sum_{ij} a_{ij}^\pi(M) e^{-M_{ij}} \leq \gamma n^2 + 2C_0.
\]

where \( a_{ij}^\pi(M) = \mathbb{P}_{X \sim H(\rho_X, M)} (A_{ij}^\pi = 1|X_{ij} = 0) \).

**Proof.** By Eq. (31), with probability \( 1 - \frac{C_0}{\gamma n^2} \), rewrite Eq. (31)

\[
\sum_{ij} \mathbb{P} (A_{ij}^\pi = 1|X) f_{ij}^\pi \leq \gamma \sum_{ij} f_{ij}^\pi.
\]

Take the expectation on the left hand side of Eq. (33), we have

\[
\mathbb{E}_{X \sim H(\rho_X, M)} \left( \sum_{ij} \mathbb{P} (A_{ij}^\pi = 1|X) f_{ij}^\pi \right)
\]

\[
= \sum_{ij} \frac{e^{-M_{ij}}}{1 + e^{-M_{ij}}} \mathbb{E} (\mathbb{P} (A_{ij}^\pi = 1, X_{ij} = 0|X)) + \mathbb{E} (\mathbb{P} (A_{ij}^\pi = 1, X_{ij} > 0|X))
\]

\[
= \sum_{ij} \frac{e^{-M_{ij}}}{1 + e^{-M_{ij}}} \mathbb{P} (A_{ij}^\pi = 1, X_{ij} = 0)
\]

\[
= \sum_{ij} \frac{1}{2} e^{-M_{ij}} \mathbb{P} (A_{ij}^\pi = 1|X_{ij} = 0) = \sum_{ij} \frac{1}{2} a_{ij}^\pi(M) e^{-M_{ij}}.
\]

where we use \( \mathbb{P} (A_{ij}^\pi = 1, X_{ij} > 0) = 0 \) and \( \mathbb{P} (X_{ij} = 0) = \frac{1}{2} + \frac{1}{2} e^{-M_{ij}} \). Take the expectation on the right hand side of Eq. (33), we have

\[
\mathbb{E}_{X \sim H(\rho_X, M)} \left( \sum_{ij} f_{ij}^\pi \right) = \gamma \sum_{ij} \left( \frac{1}{2} e^{-M_{ij}} + \frac{1}{2} (1 - e^{-M_{ij}}) \right) = \frac{\gamma n^2}{2}.
\]
Let $G(t) = t$ in the event that Eq. (33) holds; otherwise $G(t) = n^2$. Then it is easy to verify that
\[
\sum_{ij} \mathbb{P}(A_{ij}^\pi = 1|X) f_{ij}^* \leq G(\sum_{ij} f_{ij}^*)
\]
with probability 1. Take expectation on both sides, we have
\[
\sum_{ij} \frac{1}{2} a_{ij}^\pi(M) e^{-M_{ij}} \leq \frac{\gamma n^2}{2} + \frac{C_0 n^2}{n^2} \leq \frac{\gamma n^2}{2} + C_0,
\]
which completes the proof.

Next, we consider the expectation of TPR. In particular, note that
\[
\mathbb{E}_{X \sim H(p_\Lambda^*, M)} \left( \sum_{ij} \mathbb{P}(A_{ij}^\pi = 1|X) (1 - f_{ij}^*) \right)
\]
\[
= \sum_{ij} \frac{1}{1 + e^{-M_{ij}}} \mathbb{E} \left( \mathbb{P}(A_{ij}^\pi = 1, X_{ij} = 0|X) \right)
\]
\[
= \sum_{ij} \frac{1}{1 + e^{-M_{ij}}} \mathbb{P}(A_{ij}^\pi = 1, X_{ij} = 0)
\]
\[
= \sum_{ij} \frac{1}{2} \mathbb{P}(A_{ij}^\pi = 1|X_{ij} = 0) = \sum_{ij} \frac{1}{2} a_{ij}^\pi(M).
\]

(34)

Let $M^+ = 1, M^- = 1 - \frac{e^{-M^-}}{\sqrt{n}}$. For any $M \in M_n$, there are one-half $M^+$ and one-half $M^-$ entries in $M$. Note that, when observing $X_{ij} = 0$, $\pi^*$ would claim anomaly on the entry $M^+$ with priority than $M^-$, because it is more possible to observe 0 for $M^-$ in the normal situation. Indeed, we choose $\gamma$ and $p_\Lambda^*$ in a way that $\pi^*$ roughly claims anomalies for all $M^+$ entries with $X_{ij} = 0$.

Intuitively speaking, if an algorithm $\pi$ achieves the similar performance as $\pi^*$, it must be able to distinguish $M^+$ and $M^-$ from the observation $X$. However, the construction of $M_n$ prevent this distinguishability. We next provide a lemma to connect TPR and the ability of recognizing $M^+$.

**Lemma 26.** For any $\pi \in \Pi_\gamma$ and any $M \in M$,
\[e - e^{M^-} \sum_{ij} |1 \{M_{ij} = M^+\} - a_{ij}^\pi(M)| \leq \frac{n^2}{2} - \sum_{ij} a_{ij}^\pi(M) + 4C_0e.\]

**Proof.** Let
\[
x := \sum_{ij, M_{ij} = M^+} a_{ij}^\pi(M)
\]
\[
y := \sum_{ij, M_{ij} = M^-} a_{ij}^\pi(M).
\]

By Eq. (32), we have $xe^{-M^+} + ye^{-M^-} \leq \gamma n^2 + 2C_0 = \frac{n^2}{2e} + 2C_0$. Hence
\[
y \leq \left( \frac{n^2}{2e} + 2C_0 - xe^{-M^+} \right) e^{-M^-} \leq \left( \frac{n^2}{2} - x \right) e^{-M^-} + 2C_0e.\]

(35)

Furthermore,
\[
\sum_{ij} |1 \{M_{ij} = M^+\} - a_{ij}^\pi(M)| = \sum_{ij, M_{ij} = M^+} (1 - a_{ij}^\pi(M)) + \sum_{ij, M_{ij} = M^-} a_{ij}^\pi(M)
\]
\[
= \frac{n^2}{2} - x + y
\]
\[
\leq \left( \frac{n^2}{2} - x \right) + \left( \frac{n^2}{2} - x \right) \frac{e^{-M^-}}{e} + 2C_0e.\]

By Eq. (35).
Further algebra provides us
\[
\frac{n^2}{2} - x + \frac{n^2}{2} - x \frac{e^{M^*}}{e} \leq \left( \frac{n^2}{2} - x \right) \frac{e - e^{M^*}}{e} \left( e + e^{M^*} \right)
\]
\[\leq \left( \frac{n^2}{2} - x - \frac{n^2}{2} - x \frac{e^{M^*}}{e} \right) \frac{e + e^{M^*}}{e - e^{M^*}}
\]
\[= \left( \frac{n^2}{2} - x - y + \left( y - \frac{n^2}{2} - x \right) \frac{e^{M^*}}{e} \right) \frac{e + e^{M^*}}{e - e^{M^*}}
\]
\[\leq \left( \frac{n^2}{2} - x - y + 2C_0e \right) \frac{e + e^{M^*}}{e - e^{M^*}}
\]

This implies that
\[
\sum_{ij} |\mathbb{1} \{M_{ij} = M^+\} - a_{ij}^\pi(M)| + 2C_0e \leq \left( \frac{n^2}{2} - \sum_{ij} a_{ij}^\pi(M) + 2C_0e \right) \frac{e + e^{M^*}}{e - e^{M^*}}.
\]

Then, we can conclude
\[
\frac{e - e^{M^*}}{e + e^{M^*}} \sum_{ij} |\mathbb{1} \{M_{ij} = M^+\} - a_{ij}^\pi(M)| \leq \frac{n^2}{2} - \sum_{ij} a_{ij}^\pi(M) + 4C_0e
\]
which completes the proof.

Next, we show that
\[a_{ij}^\pi(M^a) \approx a_{ij}^\pi(M^b)\]
if \(M^a \approx M^b\).

**Lemma 27.** Let \(M^a \in \mathbb{R}^{n \times n}\) and \(M^b \in \mathbb{R}^{n \times n}\) only differ on two rows (WOLG, the first row and the second row). In particular, \(M_{ij}^a = \Pi_{ij}^b\) for any \(j \in [n]\) and \(i = 3, 4, \ldots, n\). Furthermore, \(M_{1j}^a = 1\) and \(M_{1j}^b = 1 - \frac{c^*}{\sqrt{n}}\) for \(j \in [n]\); \(M_{2j}^a = 1 - \frac{c^*}{\sqrt{n}}\) and \(M_{2j}^b = 1\) for \(j \in [n]\). Here \(c^* < \frac{1}{2}\). Then for any \((i, j) \in [n] \times [n],\)
\[
|a_{ij}^\pi(M^a) - a_{ij}^\pi(M^b)| \leq c^*.
\]

**Proof.** Consider some set \(S\) such that
\[
a_{ij}^\pi(M) = P_{X \sim \mathcal{H}(\rho, M)}(A_{ij} = 1 | X_{ij} = 0) = P_{X \sim \mathcal{H}(\rho, M)}(X \in S).
\]

Let \(X|Y\) be the total variation distance between \(X\) and \(Y\), \(D_{KL}(X||Y)\) be the KL-divergence between \(X\) and \(Y\). Then,
\[
|a_{ij}^\pi(M^a) - a_{ij}^\pi(M^b)| \leq \mathbb{P}_{X(M^a)}(X \in S) - \mathbb{P}_{X(M^b)}(X \in S)
\]
\[
\leq \delta(X(M^a)||X(M^b))
\]
\[
\leq \frac{1}{2}D_{KL}(X(M^a)||X(M^b))\quad \text{total variation distance}
\]
\[
\leq \sqrt{\frac{1}{2}D_{KL}(X(M^a)||X(M^b))}\quad \text{Pinke'r's inequality}
\]
\[
= \sqrt{\frac{1}{2} \sum_{ij} D_{KL}(X(M^a)_{ij}||X(M^b)_{ij})}\quad X_{ij} \text{ are independent}.
\]

Note that there are only two rows that are different between \(M^a\) and \(M^b\). Let \(X^+\) be the observation of the entry with value \(M^+\) and \(X^-\) be the observation of the entry with the value \(M^-\). Then we have
\[
\sum_{ij} D_{KL}(X(M^a)_{ij}||X(M^b)_{ij}) = nD_{KL}(X^+||X^-) + nD_{KL}(X^-||X^+).
\]
Note that $X^+ = Y^+b, X^- = Y^-b$ where $Y^+ = \text{Poisson}(M^+), Y^- = \text{Poisson}(M^-)$, and $b$ indicates whether the anomaly occurs. Hence by the data processing inequality and formula of KL-divergence of Poisson random variables,

\[
D_{KL}(X^+||X^-) \leq D_{KL}(Y^+||Y^-) = (M^+ \log(M^+/M^-) + M^- - M^+) \\
= -\log(1 - \frac{c^*}{\sqrt{n}}) - \frac{c^*}{\sqrt{n}} \\
= \frac{c^*}{\sqrt{n}} + (\frac{c^*}{n})^2 + \sum_{k=3}^{\infty} \frac{1}{k} (\frac{c^*}{\sqrt{n}})^k - \frac{c^*}{\sqrt{n}} \\
\leq \frac{(c^*)^2}{2n} + \frac{3}{3} (\frac{c^*}{\sqrt{n}})^2 \sum_{k=0}^{\infty} (\frac{c^*}{\sqrt{n}})^k \\
\leq \frac{(c^*)^2}{2n} + \frac{2c^*}{3} (\frac{c^*}{n})^2 \leq \frac{(c^*)^2}{n}.
\]

where $c^* < \frac{1}{2}$. Similarly,

\[
D_{KL}(X^-||X^+) \leq D_{KL}(Y^+||Y^-) = (M^- \log(M^-/M^+) + M^+ - M^-) \\
= (1 - \frac{c^*}{\sqrt{n}}) \log(1 - \frac{c^*}{\sqrt{n}}) + \frac{c^*}{\sqrt{n}} \\
\leq (1 - \frac{c^*}{\sqrt{n}})(-\frac{c^*}{\sqrt{n}}) + \frac{c^*}{\sqrt{n}} \\
\leq \frac{(c^*)^2}{n}.
\]

Hence,

\[
|a_{ij}^\pi(M_1) - a_{ij}^\pi(M_2)| \leq c^*.
\]

Next, we show a bound related to the “aggregated TPR” of all $M \in \mathcal{M}_n$.

**Lemma 28.**

\[
\frac{n^2}{2} - \frac{1}{2n^2} \sum_{M \in \mathcal{M}_n} \sum_{ij} a_{ij}^\pi(M) \geq -4C_0c + \frac{c^*(1-c^*)}{4} n\sqrt{n}. \tag{36}
\]

where $c$ is a constant.

**Proof.** Recall that $|\mathcal{M}_n| = \frac{1}{2^{n/2}}$. In order to use the Lemma 26, we derive a lower bound on

\[
\frac{1}{2n^2} \sum_{M \in \mathcal{M}_n} \sum_{ij} |1 \{ M_{ij} = M^+ \} - a_{ij}^\pi(M)| = \frac{1}{2n^2} \sum_{M \in \mathcal{M}_n} \sum_{ij} |1 \{ M_{ij} = M^+ \} - a_{ij}^\pi(M)|.
\]

Consider fixed $(i, j)$. Let $M^a, M^b \in \mathcal{M}_n$ be a pair of matrices such that the only different rows between $M^a, M^b$ are the $i$-th row and the $i + 1$-th row (or the $i - 1$-th row). Without loss of generality, suppose $M_{ij}^a = M^+$. Note that there are $2^{n/2 - 1}$ such pairs. Consider

\[
|1 \{ M_{ij}^a = M^+ \} - a_{ij}^\pi(M^a)| + |1 \{ M_{ij}^b = M^+ \} - a_{ij}^\pi(M^b)| \\
= |1 - a_{ij}^\pi(M^a)| + |a_{ij}^\pi(M^b)| \\
= 1 - a_{ij}^\pi(M^a) + a_{ij}^\pi(M^b) \\
\geq 1 - |a_{ij}^\pi(M^a) - a_{ij}^\pi(M^b)| \\
\geq 1 - c^*.
\]
The last inequality is due to Lemma 27. Hence,
\[
\frac{1}{2n^2} \sum_{ij} \sum_{M \in \mathcal{M}_n} |\mathbb{I} \{ M_{ij} = M^+ \} - a^\pi_{ij}(M)| \geq \frac{1}{2n^2} \sum_{ij} (2^{n/2-1})(1 - c^*) = n^2 \left(1 - \frac{c^*}{2}\right)
\]
Also note that (Recall \(c^* \leq \frac{1}{2}\))
\[
\frac{e - e^{M^+}}{e + e^{M^+}} \geq \frac{e(1 - e^{-\frac{c^*}{2}})}{e} \geq 1 - e^{-\frac{c^*}{2}} \geq \frac{c^*}{2\sqrt{n}}.
\]
Then by Lemma 26, one can obtain
\[
\frac{c^*}{2\sqrt{n}} \sum_{ij} |\mathbb{I} \{ M_{ij} = M^+ \} - a^\pi_{ij}(M)| \leq \frac{n^2}{2} - \sum_{ij} a^\pi_{ij}(M) + 4C_0e.
\]
Sum over \(M \in \mathcal{M}_n\) on both sides, we have
\[
\frac{n^2}{2} - \frac{1}{2n^2} \sum_{M \in \mathcal{M}_n} \sum_{ij} a^\pi_{ij}(M) + 4C_0e \geq \frac{c^*}{2\sqrt{n}} \frac{n^2}{2} \sum_{M \in \mathcal{M}_n} |\mathbb{I} \{ M_{ij} = M^+ \} - a^\pi_{ij}(M)| \geq \frac{c^*(1 - c^*)}{4n\sqrt{n}}.
\]
Next, we consider the ideal policy \(\pi^*(\gamma)\). Write \(\pi^*(\gamma)\) as \(\pi^*\) if there is no ambiguity.

**Lemma 29.** For any \(M \in \mathcal{M}_n\),
\[
\sum_{ij} a^\pi_{ij}(M) \geq \frac{n^2}{2} - Cn \log n - 2.
\]
where \(C\) is a constant.

**Proof.** By Lemma 14, we have, with probability \(1 - \frac{1}{n^2}\),
\[
\left|\sum_{ij} f^\pi_{ij} - n^2(1 - p^\pi)\right| \leq C_1n \log n.
\]
where \(C_1\) is a constant. Consider a policy \(\pi'\) that knows the true rate matrix \(M\). Without loss of generality, let first \(\frac{n}{2}\) rows of \(M\) be \(M^+\). Suppose \(\pi'\) claims anomalies for \((i, j)\) with \(X_{ij} = 0\) and \(i \leq (n - k_1)/2\) with \(k_1 = 8eC_1n \log n\). Then, with probability \(1 - \frac{1}{n^2}\),
\[
\sum_{ij} P\left(A^\pi_{ij} = 1\big|X\right) f^\pi_{ij} = \sum_{i \leq (n - k_1)/2} \mathbb{1}\{X_{ij} = 0\} \frac{e^{-M^+}}{1 + e^{-M^+}} \leq \frac{n(n - k_1)(1 - p^\pi)}{2e}.
\]
Then we have
\[
\sum_{ij} \mathbb{P}(A_{ij}' = 1 \mid X) f_{ij}^* \leq \frac{n^2(1 - p_{ij}^*)}{2e} - C_1 n \log n \leq \gamma \sum_{ij} f_{ij}^*.
\]

Therefore, with probability \(1 - \frac{2}{n^2}\),
\[
\sum_{ij} \mathbb{P}(A_{ij}' = 1 \mid X) (1 - f_{ij}^*) \leq \sum_{ij} \mathbb{P}(A_{ij}^* = 1 \mid X) (1 - f_{ij}^*).
\]

Finally, we have
\[
\mathbb{E}_{X(M)} \left( \sum_{ij} \mathbb{P}(A_{ij}^* = 1 \mid X) (1 - f_{ij}^*) \right) \geq \mathbb{E}_{X(M)} \left( \sum_{ij} \mathbb{P}(A_{ij}' = 1 \mid X) (1 - f_{ij}^*) \right) - \frac{n^2}{n^2}
\]
\[
= \sum_{i \leq (n - k_1) / 2} \mathbb{I} \{ X_{ij} = 0 \} \frac{1}{1 + e^{-1}} - 2
\]
\[
\geq \frac{n(n - k_1)}{4} - n \log n - 2
\]
\[
= \frac{n^2}{4} - C_n \log n - 2.
\]

**Proof of Proposition 1.** Next, we finish the proof of Proposition 1. Combining Lemma 28 and Lemma 29, we have
\[
\frac{1}{2n^2} \sum_{M \in \mathcal{M}_n} \sum_{ij} a_{ij}^*(M) - \frac{1}{2n^2} \sum_{M \in \mathcal{M}_n} \sum_{ij} a_{ij}^*(M)
\]
\[
\geq -C_n \log n - 2 - 4C_0 e + C^* (1 - C^*) \frac{1}{n \sqrt{n}}.
\]

Therefore, there exists a \(M' \in \mathcal{M}_n\), such that
\[
\sum_{ij} a_{ij}^*(M') - \sum_{ij} a_{ij}^*(M') \geq C_1 n \sqrt{n}.
\]

Finally, we have
\[
\mathbb{E}_{X \sim H(p_{ij}^*, M')} (\text{TPR}_{p_{ij}^*}(X) - \text{TPR}_{p_{ij}}(X))
\]
\[
= \mathbb{E}_{X \sim H(p_{ij}^*, M')} \left( \frac{\sum_{ij} \mathbb{P}(A_{ij}^* = 1 \mid X) - \mathbb{P}(A_{ij}^* = 1 \mid X) (1 - f_{ij}^*)}{\sum_{ij} (1 - f_{ij}^*)} \right)
\]
\[
\geq \frac{\sum_{ij} a_{ij}^*(M') - \sum_{ij} a_{ij}^*(M')} {2n^2}
\]
\[
\geq \frac{C_1}{2 \sqrt{n}}
\]
which completes the proof.

**D. Experiments**

In this section, we provide further implementation details of the experiments.

**Computing Infrastructure.** all experiments are done in a personal laptop equipped with 2.6 GHz 6-Core Intel Core i7 and 16 GB 2667 MHz DDR4. The operating system is macOS Catalina. For each instance, the running time is within seconds for our algorithm.
Near-Optimal Anomaly Detection for Matrices with Sub-Exponential Noise

Synthetic Data. We present the implementation details of our algorithm and three state-of-the-arts. For practical consideration, we implemented a slight variant of the EW algorithm where (i) the matrix completion step used the typical soft impute algorithm (Mazumder et al., 2010); (ii) the anomaly model estimation used MLE; and (iii) solving $P^{EW}$ by replacing $f_{ij}$ directly by $\frac{\hat{y}_{ij}}{x_{ij}^{a}+\hat{y}_{ij}}$. Given the observation $X_{\Omega}$, the soft impute algorithm solves the optimization problem $\min_M \| P_{\Omega}(X-M) \|^2 + \lambda \| M \|_a$ where $\lambda$ is a hyper-parameter. To tune $\lambda$, we start with a small $\lambda$ and gradually increase it until the rank of the solution fits the true rank of $M^*$ (all other algorithms also use the knowledge of the true rank). In order to generate the AUC curve for each instance, we vary $\gamma$ in our algorithm.

In the implementation of Stable-PCP, we solve the following optimization problem $(\hat{M}, \hat{A}) = \arg \min_{M,A} \| M \|_a + \lambda \| A \|_1 + \mu \| P_{\Omega}(M + A - X) \|^2_F$ by alternating optimization (Ma & Aybat, 2018). The set of anomalies is identified from $\{(i,j) \mid \hat{A}_{ij} \neq 0\}$. In order to choose suitable $(\lambda, \mu)$ and generate the AUC curve, note that when $\hat{M}$ fixed, the ratio of $\lambda/\mu$ decides the portion that will be classified as anomalies (i.e., different points on the AUC curve). Hence, we iterate the ratio $\lambda/\mu$ and then tune $\lambda$ (accordingly, $\mu$) such that the solution $\hat{M}$ fits the true rank of $M^*$. This provides an AUC curve.

In the DRMF algorithm, we implement the Algorithm 1 in (Xiong et al., 2011)\(^7\) to solve the following optimization problem $(\hat{M}, \hat{A}) = \arg \min_{M,A} \| P_{\Omega}(X - A - M) \|_F$ with the constraints $\text{rank}(M) \leq r, \| A \|_0 \leq e$. The set of anomalies is identified from $\{(i,j) \mid \hat{A}_{ij} \neq 0\}$. Here, we provide the true rank $r$ and vary $e$ for the DRMF algorithm to generate the AUC curve.

For the RMC algorithm (Klopp et al., 2017), the authors propose the following optimization problem $(\hat{M}, \hat{A}) = \arg \min_{M,A} \| M \|_a + \lambda \| A \|_1 + \mu \| P_{\Omega}(M + A - X) \|^2_F$ with constraints $\| M \|_{\text{max}} \leq a, \| A \|_{\text{max}} \leq a$. This is effectively the Stable-PCP algorithm with the max norm constraints. We choose $a = k \| M^* \|_{\text{max}}$ for some constant scale $k > 1$. Then we implement RMC based on Stable-PCP and a projection of $(M, A)$ into the set with max norm constraints in every iteration during the alternating optimization.

We also study the limitation of our algorithm, in which the performance starts to degrade. Figure 3 shows that the problem instances (in the experiment of the synthetic data) where the AUC of EW was furthest away from the ideal AUC (20th percentile). The results show largely intuitive characteristics: higher $\alpha^*$ (so anomalies look similar to non-anomalous entries), lower $p_0$, higher $p^A_\lambda$ and higher $r$ (so that $M^*$ is harder to estimate). The behavior with respect to $\hat{M}^*$ is surprising but was consistently observed across other ensembles as well.

Real Data. We estimate the rank of $M^*$ to obtain $r \sim 30$ via cross-validation. The observation $X$ is generated

---

\(^7\)Although (Xiong et al., 2011) does not consider the partial observation scenario, but the generalization to address missing entries is straightforward.
in a same way described in the synthetic data. The EW and Stable-PCP algorithms are implemented in a same way as in the synthetic data with rank information $r$. 