# Statistical Analysis of Multi-Relational Network Recovery

**Zhi Wang [1], Xueying Tang [2,*] and Jingchen Liu [1]**

[1] *Department of Statistics, Columbia University*
[2] *Department of Mathematics, University of Arizona*

Correspondence*:
Xueying Tang
Address: 617 N. Santa Rita Ave., Tucson, AZ 85721, USA.
Email: xytang@math.arizona.edu

## ABSTRACT

In this paper, we develop asymptotic theories for a class of latent variable models for large-scale multi-relational networks. In particular, we establish consistency results and asymptotic error bounds for the (penalized) maximum likelihood estimators when the size of the network tends to infinity. The basic technique is to develop a non-asymptotic error bound for the maximum likelihood estimators through large deviations analysis of random fields. We also show that these estimators are nearly optimal in terms of minimax risk.

**Keywords: multi-relational network, knowledge graph completion, tail probability, risk, asymptotic analysis, non-asymptotic analysis, maximum likelihood estimation**

## 1 INTRODUCTION

A multi-relational network (MRN) describes multiple relations among a set of entities simultaneously. Our work on MRNs is mainly motivated by its applications to knowledge bases that are repositories of information. Examples of knowledge bases include WordNet [1], Unified Medical Language System [2], and Google Knowledge Graph (https://developers.google.com/knowledge-graph). They have been used as the information source in many natural language processing tasks such as word-sense disambiguation and machine translation [3, 4, 5]. A knowledge base often includes knowledge on a large number of real-world objects or concepts. When a knowledge base is characterized by MRN, the objects and concepts corresponds to nodes, and knowledge types are relations. Figure 1 provides an excerpt from an MRN in which "Earth", "Sun" and "solar system" are three nodes. The knowledge about the orbiting patterns of celestial objects forms a relation "orbit", and the knowledge on classification of the objects forms another relation "belong to" in the MRN.

An important task of network analysis is to recover the unobserved network based on data. In this paper, we consider a latent variable model for MRNs. The presence of an edge from node $i$ to node $j$ of relation type $k$ is a Bernoulli random variable $Y_{ijk}$ with success probability $M_{ijk}$. Each node is associated with a vector, $\boldsymbol{\theta}$, called the embedding of the node. The probability $M_{ijk}$ is modeled as a function $f$ of the embeddings, $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$, and a relation-specific parameter vector $\boldsymbol{w}_k$. This is a natural generalization of the latent space model for single-relational networks [6]. Recently, it has been successfully applied to knowledge base analysis [7, 8, 9, 10, 11, 12, 13, 14]. Various forms of $f$ are proposed such as distance models [7], bilinear models [12, 13, 14], and neural networks [15]. Computational algorithms are proposed

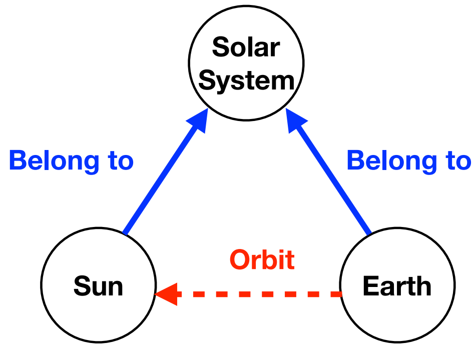


**Figure 1.** An example of the MRN representation of a knowledge base.

to improve link prediction for knowledge bases [16, 17]. The statistical properties of the embedding-based MRN models have not been rigorously studied. It remains unknown whether and to what extent the underlying distribution of MRN can be recovered, especially when there are a large number of nodes and relations.

The results in this paper fill in the void by studying the error bounds and asymptotic behaviors of the estimators for $M_{ijk}$'s for a general class of models. This is a challenging problem due to the following facts. Traditional statistical inference of latent variable models often requires a (proper or improper) prior distribution for $\boldsymbol{\theta}_i$. In such settings, one works with the marginalized likelihood with $\boldsymbol{\theta}_i$ integrated out. For the analysis of MRN, the sample size and the latent dimensions are often so large that the above-mentioned inference approaches are computationally infeasible. For instance, a small-scale MRN could have a sample size as large as a few million, and the dimension of the embeddings is as large as several hundred. Therefore, in practice, the prior distribution is often dropped, and the latent variables $\boldsymbol{\theta}_i$'s are considered as additional parameters and estimated via maximizing the likelihood or penalized likelihood functions. The parameter space is thus substantially enlarged due to the addition of $\boldsymbol{\theta}_i$'s whose dimension is proportionate to the number of entities. As a result, in the asymptotic analysis, we face a double-asymptotic regime of both the sample size and the parameter dimension.

In this paper, we develop results for the (penalized) maximum likelihood estimator of such models and show that under regularity conditions the estimator is consistent. In particular, we overcome the difficulty induced by the double-asymptotic regime via non-asymptotic bounds for the error probabilities. Then, we show that the distribution of MRN can be consistently estimated in terms of average Kullback-Leibler (KL) divergence even when the latent dimension increases slowly as the sample size tends to infinity. A probability error bound is also provided together with the upper bound for the risk (expected KL divergence). We further study the lower bound and show the near-optimality of the estimator in terms of minimax risk. Besides the average KL divergence, similar results can be established for other criteria such as link prediction accuracy.

The outline of the remaining sections is as follows. In Section 2, we provide the model speicification and formulate the problem. Our main results are presented in Section 3. Finite sample performance is examined in Section 4 through simulated and real data examples. Concluding remarks are included in Section 5.

## 2 PROBLEM SETUP

### 2.1 Notation

Let $|\cdot|$ be the cardinality of a set and $\times$ be the Cartesian product. Set $\{1,\ldots,N\}$ is denoted by $[N]$. The sign function $\mathrm{sgn}(x)$ is defined to be 1 for $x \geq 0$ and 0 otherwise. The logistic function is denoted by $\sigma(x) = e^x/(1+e^x)$. Let $1_A$ be the indicator function on event $A$. We use $U[a,b]$ to denote the uniform distribution on $[a,b]$ and $\mathrm{Ber}(p)$ to denote the Bernoulli distribution with probability $p$. The KL divergence between $\mathrm{Ber}(p)$ and $\mathrm{Ber}(q)$ is written as $D(p||q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$. We use $\|\cdot\|$ to denote the Euclidean norm for vectors and the Frobenius norm for matrices.

For two real positive sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n = O(b_n)$ if $\limsup_{n\to\infty} a_n/b_n < \infty$. Similarly, we write $a_n = \Omega(b_n)$ if $\limsup_{n\to\infty} b_n/a_n < \infty$ and $a_n = o(b_n)$ if $\lim_{n\to\infty} a_n/b_n = 0$. We denote $a_n \lesssim b_n$ if $\limsup_{n\to\infty} a_n/b_n \leq 1$. When $\{a_n\}$ and $\{b_n\}$ are negative sequences, $a_n \lesssim b_n$ means $\liminf_{n\to\infty} a_n/b_n \geq 1$. In some places, we use $b_n \gtrsim a_n$ as an interchangeable notation of $a_n \lesssim b_n$. Finally, if $\lim_{n\to\infty} a_n/b_n = 1$, we write $a_n \sim b_n$.

### 2.2 Model

Consider an MRN with $N$ entities and $K$ relations. Given $i,j \in [N]$ and $k \in [K]$, the triple $\lambda = (i,j,k)$ corresponds to the edge from entity $i$ to entity $j$ of relation $k$. Let $\Lambda = [N]\times[N]\times[K]$ denote the set of all edges. We assume in this paper that an edge can be either present or absent in a network and use $Y_\lambda \in \{0,1\}$ to indicate the presence of edge $\lambda$. In some scenarios, the status of an edge may have more than two types. Our analysis can be generalized to accommodate these cases.

We associate each entity $i$ with a vector $\boldsymbol{\theta}_i$ of dimension $d_E$ and each relation $k$ with a vector $\boldsymbol{w}_k$ of dimension $d_R$. Let $\mathcal{E} \subseteq \mathbb{R}^{d_E}$ be a compact domain where the embeddings $\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_N$ live. We call $\mathcal{E}$ the entity space. Similarly, we define a compact relation space $\mathcal{R} \subseteq \mathbb{R}^{d_R}$ for the relation-specific parameters $\boldsymbol{w}_1,\ldots,\boldsymbol{w}_K$. Let $\boldsymbol{x} = (\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_N,\boldsymbol{w}_1,\ldots,\boldsymbol{w}_K)$ be a vector in the product space $\Theta = \mathcal{E}^N \times \mathcal{R}^K$. The parameters associated with edge $\lambda = (i,j,k)$ is then $\boldsymbol{x}_\lambda = (\boldsymbol{\theta}_i,\boldsymbol{\theta}_j,\boldsymbol{w}_k)$. We assume that given $\boldsymbol{x}$, elements in $\{Y_\lambda \mid \lambda \in \Lambda\}$ are independent with each other and that the log odds of $Y_\lambda = 1$ is

$$\log\frac{P(Y_\lambda = 1|\boldsymbol{x})}{P(Y_\lambda = 0|\boldsymbol{x})} = \phi(\boldsymbol{x}_\lambda), \text{ for } \lambda \in \Lambda. \tag{1}$$

Here $\phi$ is defined on $\mathcal{E}^2 \times \mathcal{R}$, and $\phi(\boldsymbol{x}_\lambda)$ is often called the score of edge $\lambda$.

We will use $Y$ to represent the $N\times N\times K$ tensor formed by $\{Y_\lambda \mid \lambda \in \Lambda\}$ and $M(\boldsymbol{x})$ to represent the corresponding probability tensor $\{P(Y_\lambda = 1 \mid \boldsymbol{x}) \mid \lambda \in \Lambda\}$. Our model is given by

$$Y_\lambda \sim \mathrm{Ber}(M_\lambda(\boldsymbol{x}^*)), \tag{2}$$

$$M_\lambda(\boldsymbol{x}) = \sigma(\phi(\boldsymbol{x}_\lambda)), \lambda \in \Lambda, \tag{3}$$

where $\boldsymbol{x}^*$ stands for the true value of $\boldsymbol{x}$ and $Y_\lambda$'s are independent. In the above model, the probability of the presence of an edge is entirely determined by the embeddings of the corresponding entities and the relation-specific parameters. This imposes a low-dimensional latent structure on the probability tensor $M^* = M(\boldsymbol{x}^*)$.

We specify our model using a generic function $\phi$. It includes various existing models as special cases. Below are two examples of $\phi$.

1.Distance model [7].

$$\phi(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j, \boldsymbol{w}_k) = b_k - \|\boldsymbol{\theta}_i + \boldsymbol{a}_k - \boldsymbol{\theta}_j\|^2, \tag{4}$$

where $\boldsymbol{\theta}_i, \boldsymbol{\theta}_j, \boldsymbol{a}_k \in \mathbb{R}^d$, $b_k \in \mathbb{R}$ and $\boldsymbol{w}_k = (\boldsymbol{a}_k, b_k)$. In the distance model, relation $k$ from node $i$ to node $j$ is more likely to exist if $\boldsymbol{\theta}_i$ shifted by $\boldsymbol{a}_k$ is closer to $\boldsymbol{\theta}_j$ under the Euclidean norm.

2.Bilinear model [9].

$$\phi(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j, \boldsymbol{w}_k) = \boldsymbol{\theta}_i^T \operatorname{diag}(\boldsymbol{w}_k)\boldsymbol{\theta}_j, \tag{5}$$

where $\boldsymbol{\theta}_i, \boldsymbol{\theta}_j, \boldsymbol{w}_k \in \mathbb{R}^d$ and $\operatorname{diag}(\boldsymbol{w}_k)$ is a diagonal matrix with $\boldsymbol{w}_k$ as the diagonal elements. Model (5) is a special case of the more general model $\phi(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j, \boldsymbol{w}_k) = \boldsymbol{\theta}_i^T W_k \boldsymbol{\theta}_j$, where $W_k \in \mathbb{R}^{d\times d}$ is a matrix parametrized by $\boldsymbol{w}_k \in \mathbb{R}^{d_R}$. Trouillon et al. [12], Nickel et al. [13] and Liu et al. [14] explored different ways of constructing $W_k$.

Very often, only a small portion of the network is observed [18]. We assume that each edge in the MRN is observed independently with probability $\gamma$ and that the observation of an edge is independent of $Y$. Let $\mathcal{S} \subset \Lambda$ be the set of observed edges. Then the elements in $\mathcal{S}$ are independent draws from $\Lambda$. For convenience, we use $n$ to represent the expected number of observed edges, namely, $n = E[|\mathcal{S}|] = \gamma|\Lambda| = \gamma N^2 K$. Our goal is to recover the underlying probability tensor $M^*$ based on the observed edges $\{Y_\lambda \mid \lambda \in \mathcal{S}\}$.

REMARK 1. *Ideally, if there exists $\boldsymbol{x}^*$ such that $Y_\lambda = sgn\left(M_\lambda(\boldsymbol{x}^*) - \frac{1}{2}\right)$ for all $\lambda \in \Lambda$, then $Y$ can be recovered with no error under $\boldsymbol{x}^*$. This is, however, a rare case in practice, especially for large-scale MRN. A relaxed assumption is that $Y$ can be recovered with some low dimensional $\boldsymbol{x}^*$ and noise $\{\epsilon_\lambda\}$ such that*

$$Y_\lambda = sgn\left(M_\lambda(\boldsymbol{x}^*) + \epsilon_\lambda - \frac{1}{2}\right), \quad \epsilon_\lambda \overset{i.i.d}{\sim} U\left[-\frac{1}{2}, \frac{1}{2}\right], \quad \forall \lambda \in \Lambda. \tag{6}$$

*By introducing the noise term, we formulate the deterministic MRN as a random graph. The model described in (2) is an equivalent but simpler form of (6).*

## 2.3 Estimation

According to (2), the log-likelihood function of our model is

$$l(\boldsymbol{x}; Y_{\mathcal{S}}) = \sum_{\lambda \in \mathcal{S}} Y_\lambda \log M_\lambda(\boldsymbol{x}) + (1 - Y_\lambda)\log(1 - M_\lambda(\boldsymbol{x})). \tag{7}$$

We omit the terms $\sum_{\lambda\in\mathcal{S}} \log \gamma + \sum_{\lambda\notin\mathcal{S}} \log(1-\gamma)$ in (7) since $\gamma$ is not the parameter of interest. To obtain an estimator of $M^*$, we take the following steps.

1. Obtain the maximum likelihood estimator (MLE) of $\boldsymbol{x}^*$,

$$\hat{\boldsymbol{x}} = \underset{\boldsymbol{x}\in\Theta}{\operatorname{argmax}}\, l(\boldsymbol{x}; Y_{\mathcal{S}}). \tag{8}$$

2. Use the plug-in estimator

$$\hat{M} = M(\hat{\boldsymbol{x}}) \tag{9}$$

as an estimator of $M^*$.

In (8), the estimator $\hat{\boldsymbol{x}}$ is a maximizer over the compact parameter space $\Theta = \mathcal{E}^N \times \mathcal{R}^K$. The dimension of $\Theta$ is

$$m = Nd_E + Kd_R,$$

which grows linearly in the number of entities $N$ and the number of relations $K$.

### 2.4 Evaluation criteria

We consider the following criteria to measure the error of the above-mentioned estimator. They will be used in both the main results and numerical studies.

(a) Average KL divergence of the predictive distribution from the true distribution

$$L(\hat{M}, M^*) = \frac{1}{|\Lambda|} \sum_{\lambda \in \Lambda} D(M^*_\lambda || \hat{M}_\lambda). \tag{10}$$

(b) Mean squared error of the predicted scores

$$MSE_\phi = \frac{1}{|\Lambda|} \sum_{\lambda \in \Lambda} \left(\phi(\hat{\boldsymbol{x}}_\lambda) - \phi(\boldsymbol{x}^*_\lambda)\right)^2. \tag{11}$$

(c) Link prediction error

$$\widehat{err} = \frac{1}{|\Lambda|} \sum_{\lambda \in \Lambda} 1_{\hat{Y}_\lambda \neq Y^*_\lambda}, \tag{12}$$

where $\hat{Y}_\lambda = \text{sgn}\left(\hat{M}_\lambda - \frac{1}{2}\right)$ and $Y^*_\lambda = \text{sgn}\left(M^*_\lambda - \frac{1}{2}\right)$.

REMARK 2. *The latent attributes of entities and relations are often not identifiable, so the MLE $\hat{\boldsymbol{x}}$ is not unique. For instance, in (4), the values of $\phi$ and $M(\boldsymbol{x})$ remain the same if we replace $\boldsymbol{\theta}_i$ and $\boldsymbol{a}_k$ respectively by $\Gamma\boldsymbol{\theta}_i + \boldsymbol{t}$ and $\Gamma\boldsymbol{a}_k$, where $\boldsymbol{t}$ is an arbitrary vector in $\mathbb{R}^{d_E}$ and $\Gamma$ is an orthonormal matrix. Therefore, we consider the mean squared error of scores, which are identifiable.*

## 3 MAIN RESULTS

We first provide results of the MLE in terms of KL divergence between the estimated and the true model. Specifically, we investigate the tail probability $P(L(\hat{M}, M^*) > t)$ and the expected loss $E[L(\hat{M}, M^*)]$. In Section 3.1, we discuss upper bounds for the two quantities. The lower bounds are provided in Section 3.2. In Section 3.3, we extend the results to penalized maximum likelihood estimators (pMLE) and other loss functions. All proofs are deferred to the Appendix.

### 3.1 Upper bounds

We first present an upper bound for the tail probability $P(L(\hat{M}, M^*) > t)$ in Lemma 1. The result depends on the tensor size, the number of observed edges, the functional form of $\phi$, and the geometry of parameter space $\Theta$. The lemma explicitly quantifying the impact of these element on the error probability. It is key to the subsequent analyses. Lemma 2 gives a non-asymptotic upper bound for the expected loss (risk). We then establish the consistency of $\hat{M}$ and the asymptotic error bounds in Theorem 1.

We will make the following assumptions throughout this section.

ASSUMPTION 1. *$\boldsymbol{x}^* \in \Theta = \mathcal{E}^N \times \mathcal{R}^K$, where $\mathcal{E}$ and $\mathcal{R}$ are Euclidean balls of radius $U$.*

ASSUMPTION 2. *The function $\phi$ is Lipschitz continuous under the Euclidean norm,*

$$|\phi(\boldsymbol{u}) - \phi(\boldsymbol{v})| \leq \alpha \|\boldsymbol{u} - \boldsymbol{v}\|, \quad \forall \boldsymbol{u}, \boldsymbol{v} \in \mathcal{E}^2 \times \mathcal{R}, \tag{13}$$

*where $\alpha$ is a Lipschitz constant.*

Assumption 1 is imposed for technical convenience. The results can be easily extended to general compact parameter spaces. Let $C = \sup_{\boldsymbol{u} \in \mathcal{E}^2 \times \mathcal{R}} |\phi(\boldsymbol{u})|$. Without loss of generality, we assume that $C \geq 2$.

LEMMA 1. *Consider $\hat{M}$ defined in (9) and the average KL divergence $L$ in (10). Under Assumptions 1 and 2, for every $t > 0$, $\beta > 0$ and $0 < s < nt$,*

$$P\left(L(\hat{M}, M^*) \geq t\right) \leq \exp\left\{-\frac{nt - s}{C} h\left(\frac{1}{2} - \frac{s}{2nt}\right)\right\} \left(1 + \frac{2\sqrt{3}\alpha U n(1+\beta)}{s}\right)^m + \exp\left\{-n\beta h(\beta)\right\}, \tag{14}$$

*where $m = Nd_E + Kd_R$ is the dimension of $\Theta$, $n = \gamma N^2 K$ is the expected number of observations, and $h(u) = (1 + \frac{1}{u})\log(1+u) - 1$.*

In the proof of Lemma 1, we use Bennett's inequality to develop a uniform bound that does not depend on the true parameters. It is sufficient for the current analysis. If the readers need sharper bounds, they can read through the proof and replace the Bennett's bound by the usual large deviation rate function which provides a sharp exponential bound that depends on the true parameters. We don't pursue this direction in this paper.

Lemma 2 below gives an upper bound of risk $E[L(\hat{M}, M^*)]$, which follows from Lemma 1.

LEMMA 2. *Consider $\hat{M}$ defined in (9) and loss function $L$ in (10). Let $C_1 = 18C$, $C_2 = 8\sqrt{3}\alpha U$ and $C_3 = 2\max\{C_1, C_2\}$. If Assumptions 1 and 2 hold and $\frac{n}{m} \geq C_2 + e$, then*

$$E[L(\hat{M}, M^*)] \leq C_3 \frac{m}{n} \log\frac{n}{m} + \frac{C_1}{n}\exp\left\{-m\log\frac{n}{m}\right\} + \frac{3}{n}\exp\left\{-\frac{1}{3}\left(n + C_3 m \log\frac{n}{m}\right)\right\}. \tag{15}$$

We are interested in the asymptotic behavior of the tail probability in two scenarios: (i) $t$ is a fixed constant and (ii) $t$ decays to zero as the number of entities $N$ tends to infinity. The following theorem gives an asymptotic upper bound for the tail probability and the risk.

THEOREM 1. *Consider $\hat{M}$ defined in (9) and the loss function $L$ in (10). Let the number of entities $N \to \infty$ and $C, K, U, d_E, d_R, \alpha$, and $\gamma$ be fixed constants. If Assumptions 1 and 2 hold, we have the following asymptotic inequalities.*

*When $t$ is a fixed constant,*

$$\log P(L(\hat{M}, M^*) \geq t) \lesssim -\frac{t}{5C} n. \tag{16}$$

*When $t = 10C\frac{m}{n}\log\frac{n}{m}$,*

$$\log P(L(\hat{M}, M^*) \geq t) \lesssim -m\log\frac{n}{m}. \tag{17}$$

*Furthermore,*

$$E[L(\hat{M}, M^*)] \lesssim 10C\frac{m}{n}\log\frac{n}{m}. \tag{18}$$

The consistency of $\hat{M}$ is implied by (16) and the rate of convergence is $|\log P(L(\hat{M}, M^*) \geq t)| = \Omega(N^2)$ if $t$ is a fixed constant. The rate decreases to $\Omega(N \log N)$ for the choice of $t$ producing (17). It is also implied by (17) that $L(\hat{M}, M^*) = O(\frac{1}{N}\log N)$ with high probability. We show in the next section that this upper bound is reasonably sharp.

The condition that $K, U, d_E, d_R$, and $\alpha$ are fixed constants can be relaxed. For instance, we can let $U$, $d_E$, $d_R$, and $\alpha$ go to infinity slowly at the rate $O(\log N)$ and $K$ at the rate $O(N)$. We can let $\gamma$ go to zero provided that $\frac{m}{n}\log\frac{n}{m} = o(1)$.

### 3.2 Lower bounds

We show in Theorem 2 that the order of the minimax risk is $\Omega(\frac{m}{n})$, which implies the near optimality of $\hat{M}$ in (9) and the upper bound $O(\frac{m}{n}\log\frac{n}{m})$ in Theorem 1. To begin with, we introduce the following definition and assumption.

DEFINITION 1. *For $\boldsymbol{u} = (\boldsymbol{\theta}, \boldsymbol{\theta}', \boldsymbol{w}) \in \mathcal{E}^2 \times \mathcal{R}$, the $r$-neighborhood of $\boldsymbol{u}$ is*

$$\mathcal{N}_r(\boldsymbol{u}) = \left\{(\boldsymbol{\eta}, \boldsymbol{\eta}', \boldsymbol{\zeta}) \in \mathcal{E}^2 \times \mathcal{R} \mid \|\boldsymbol{\eta} - \boldsymbol{\theta}\| \leq r, \|\boldsymbol{\eta}' - \boldsymbol{\theta}'\| \leq r, \|\boldsymbol{\zeta} - \boldsymbol{w}\| \leq r\right\}.$$

*Similarly, for $\boldsymbol{x} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N, \boldsymbol{w}_1, \ldots, \boldsymbol{w}_K) \in \mathcal{E}^N \times \mathcal{R}^K$, the $r$-neighborhood of $\boldsymbol{x}$ is*

$$\mathcal{N}_r(\boldsymbol{x}) = \left\{(\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_N, \boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_K) \in \mathcal{E}^N \times \mathcal{R}^K \mid \|\boldsymbol{\eta}_i - \boldsymbol{\theta}_i\| \leq r, \|\boldsymbol{\zeta}_k - \boldsymbol{w}_k\| \leq r, \forall i \in [N], k \in [K]\right\}.$$

ASSUMPTION 3. *There exists $\boldsymbol{u}_0 \in \mathcal{E}^2 \times \mathcal{R}$ and $r, \kappa > 0$ such that $\mathcal{N}_r(\boldsymbol{u}_0) \subset \mathcal{E}^2 \times \mathcal{R}$ and*

$$|\sigma(\phi(\boldsymbol{u})) - \sigma(\phi(\boldsymbol{v}))| \geq \kappa\|\boldsymbol{u} - \boldsymbol{v}\|, \quad \forall \boldsymbol{u}, \boldsymbol{v} \in \mathcal{N}_r(\boldsymbol{u}_0). \tag{19}$$

THEOREM 2. *Let $b = \sup_{\boldsymbol{u} \in \mathcal{N}_r(\boldsymbol{u}_0)} \sigma(\phi(\boldsymbol{u}))$. Under Assumptions 2 and 3, if $r^2 \geq \frac{(m/16-1)b(1-b)}{12\alpha^2 n}$, then for any estimator $\hat{M}$, there exists $\boldsymbol{x}^* \in \Theta$ such that*

$$P\left(L(\hat{M}, M^*) > \tilde{C}\frac{m/16 - 1}{n}\right) \geq \frac{1}{2}, \tag{20}$$

*where $\tilde{C} = \frac{\kappa^2 b(1-b)}{108\alpha^2}$. Consequently, the minimax risk*

$$\min_{\hat{M}} \max_{M^*} E[L(\hat{M}, M^*)] \geq \tilde{C}\frac{m/16 - 1}{2n}. \tag{21}$$

### 3.3 Extensions

#### 3.3.1 Reguralization

In this section, we extend our asymptotic results in Theorem 1 to regularized estimators. In practice, regularization is often considered to prevent overfitting. We consider a regularization similar to elastic net

[19]

$$l_\rho\left(\boldsymbol{x}; Y_{\mathcal{S}}\right) = l(\boldsymbol{x}; Y_{\mathcal{S}}) - \rho_1 \|\boldsymbol{x}\|_1 - \rho_2 \|\boldsymbol{x}\|^2, \tag{22}$$

where $\|\cdot\|_1$ stands for $L_1$ norm and $\rho_1, \rho_2 \geq 0$ are regularization parameters. The pMLE is

$$\hat{\boldsymbol{x}} = \underset{\boldsymbol{x}\in\Theta}{\operatorname{argmax}}\, l_\rho(\boldsymbol{x}; Y_{\mathcal{S}}). \tag{23}$$

Note that the MLE in (8) is a special case of the pMLE above with $\rho_1 = \rho_2 = 0$. Since $\hat{\boldsymbol{x}}$ is shrunk towards $\mathbf{0}$, without loss of generality, we assume that $\mathcal{E}$ and $\mathcal{R}$ are centered at $\mathbf{0}$. We generalize Theorem 1 to pMLE in the following theorem.

THEOREM 3. *Consider the estimator $\hat{M}$ given by (23) and (9) and the loss function $L$ in (10). Let the number of entities $N \to \infty$ and $C, K, U, d_E, d_R, \alpha, \gamma$ be absolute constants. If Assumptions 1 and 2 hold and $\rho_1 + \rho_2 = o(\log N)$, then asymptotic inequalities (16), (17), and (18) in Theorem 1 hold.*

### 3.3.2 Other loss functions

We present some results for the mean squared error loss $MSE_\phi$ defined in (11) and the link prediction error $\widehat{err}$ defined in (12). Corollaries 1 and 2 give upper and lower bounds for $MSE_\phi$, and Corollary 3 gives an upper bound for $\widehat{err}$ under an additional assumption.

COROLLARY 1. *Under the setting of Theorem 3 with the loss function replaced by $MSE_\phi$, we have the following asymptotic results.*
*If $t$ is a fixed constant,*

$$\log P\left(MSE_\phi \geq t\right) \lesssim -\frac{5\sigma(C)\left(1-\sigma(C)\right)t}{2C} n. \tag{24}$$

*If $t = \frac{20C}{\sigma(C)(1-\sigma(C))} \frac{m}{n} \log \frac{n}{m}$,*

$$\log P\left(MSE_\phi \geq t\right) \lesssim -m \log \frac{n}{m}. \tag{25}$$

*Furthermore,*

$$E\left[MSE_\phi\right] \lesssim \frac{20C}{\sigma(C)\left(1-\sigma(C)\right)} \frac{m}{n} \log \frac{n}{m}. \tag{26}$$

COROLLARY 2. *Under the setting of Theorem 2 with the loss function replaced by $MSE_\phi$, we have*

$$P\left(MSE_\phi > \tilde{C}\frac{m/16-1}{8n}\right) \geq \frac{1}{2}, \tag{27}$$

*and*

$$\min_{\hat{M}} \max_{M^*} E\left[MSE_\phi\right] \geq \tilde{C}\frac{m/16-1}{16n}. \tag{28}$$

ASSUMPTION 4. *There exists $\varepsilon > 0$ such that $\left|M^*_\lambda - \frac{1}{2}\right| \geq \varepsilon$ for every $\lambda \in \Lambda$.*

COROLLARY 3. *Under the setting of Theorem 3 with the loss function replaced by $\widehat{err}$ and Assumption 4 added, we have the following asymptotic results.*
*If $t$ is a fixed constant,*

$$\log P\left(\widehat{err} \geq t\right) \lesssim -\frac{2\varepsilon^2 t}{5C} n. \tag{29}$$

*If* $t = \frac{5C}{\varepsilon^2}\frac{m}{n}\log\frac{n}{m}$,

$$\log P\left(\widehat{err} \geq t\right) \lesssim -m\log\frac{n}{m}. \tag{30}$$

*Furthermore,*

$$E\left[\widehat{err}\right] \lesssim \frac{5C}{\varepsilon^2}\frac{m}{n}\log\frac{n}{m}. \tag{31}$$

### 3.3.3 Sparse representations

We are interested in sparse entity embeddings and relation parameters. Let $\|\cdot\|_0$ be the number of non-zero elements of a vector and $\tau$ be a prespecified sparsity level of $\boldsymbol{x}$ (i.e. the proportion of nonzero elements). Let $m_\tau = m\tau$ be the upper bound of non-zero parameters, that is, $\|\boldsymbol{x}^*\|_0 \leq m_\tau$. Consider the following estimator

$$\hat{\boldsymbol{x}} = \underset{\boldsymbol{x}\in\Theta}{\operatorname{argmax}}\, l\left(\boldsymbol{x};\boldsymbol{Y}_{\mathcal{S}}\right) \quad \text{subject to} \quad \|\boldsymbol{x}\|_0 \leq m_\tau. \tag{32}$$

The estimator defined above maximizes the $L_0$-penalized log-likelihood.

THEOREM 4. *Consider $\hat{M}$ defined in (32) and (9) and the loss function $L$ in (10). Let the number of entities $N \to \infty$ and $\tau, C, K, U, d_E, d_R, \alpha$ be absolute constants. Under Assumptions 1 and 2, the following asymptotic inequalities hold.*

*If $t$ is a fixed constant,*

$$\log P(L(\hat{M}, M^*) \geq t) \lesssim -\frac{t}{5C}n. \tag{33}$$

*If* $t = 10C\frac{m_\tau}{n}\log\frac{n}{m_\tau}$,

$$\log P(L(\hat{M}, M^*) \geq t) \lesssim -m_\tau\log\frac{n}{m_\tau}. \tag{34}$$

*Furthermore,*

$$E[L(\hat{M}, M^*)] \lesssim 10C\frac{m_\tau}{n}\log\frac{n}{m_\tau}. \tag{35}$$

We omit the results for other loss functions as well as the lower bounds since they can be analogously obtained.

# 4 NUMERICAL EXAMPLES

In this section, we demonstrate the finite sample performance of $\hat{M}$ through simulated and real data examples. Throughout the numerical experiments, AdaGrad algorithm [20] is used to compute $\hat{\boldsymbol{x}}$ in (8) or (23). It is a first-order optimization method that combines stochastic gradient descent (SGD) [21] with adaptive step sizes for finding the local optima. Since the objective function in (8) is non-convex, a global maximizer is not guaranteed. Our objective function usually has many global maximizers, but, empirically, we found the algorithm works well on MRN recovery and the recovery performance is insensitive to the choice of the starting point of SGD. Computationally, SGD is also more appealing to handle large-scale MRNs than those more expensive global optimization methods.

## 4.1 Simulated Examples

In the simulated examples, we fix $K = 20$, $d_E = 20$ and consider various choices of $N$ ranging from 100 to 10,000 to investigate the estimation performance as $N$ grows. The function $\phi$ we consider is a combination of the distance model (4) and the bilinear model (5),

$$\phi(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j, \boldsymbol{w}_k) = (\boldsymbol{\theta}_i + \boldsymbol{a}_k - \boldsymbol{\theta}_j)^T \operatorname{diag}(\boldsymbol{b}_k)(\boldsymbol{\theta}_i + \boldsymbol{a}_k - \boldsymbol{\theta}_j), \tag{36}$$

where $\boldsymbol{\theta}_i, \boldsymbol{\theta}_j, \boldsymbol{a}_k, \boldsymbol{b}_k \in \mathbb{R}^d$ and $\boldsymbol{w}_k = (\boldsymbol{a}_k, \boldsymbol{b}_k)$. We independently generate the elements of $\boldsymbol{\theta}_i^*$, $\boldsymbol{a}_k^*$, and $\boldsymbol{b}_k^*$ from normal distributions $N(0,1)$, $N(0,1)$, and $N(0, 0.25)$, respectively, where $N(\mu, \sigma^2)$ denotes the normal distribution with mean $\mu$ and variance $\sigma^2$. To guarantee that the parameters are from a compact set, the normal distributions are truncated to the interval [-20, 20]. Given the latent attributes, each $Y_{ijk}$ is generated from the Bernoulli distribution with success probability $M_{ijk}^* = \sigma(\phi(\boldsymbol{\theta}_i^*, \boldsymbol{\theta}_j^*, \boldsymbol{w}_k^*))$. The observation probability $\gamma$ takes value from $\{0.005, 0.01, 0.02\}$. For each combination of $\gamma$ and $N$, 100 independent datasets are generated. For each dataset, we compute $\hat{\boldsymbol{x}}$ and $\hat{M}$ in (8) and (9) with AdaGrad algorithm and then calculate $L(\hat{M}, M^*)$ defined in (10) as well as the link prediction error $\widehat{err}$ defined in (12). The two types of losses are averaged over the 100 datasets for each combination of $N$ and $\gamma$ to approximate the theoretical risks $E[L(\hat{M}, M^*)]$ and $E[\widehat{err}]$. These quantities are plotted against $N$ in log scale in Figure 2. As the figure shows, in general, both risks decrease as $N$ increases. When $N$ is small, $n/m$ is not large enough to satisfy the condition $n/m \geq C_2 + e$ in Lemma 2 and the expected KL risk increases at the beginning. After $N$ gets sufficiently large, the trend agrees with our asymptotic analysis.

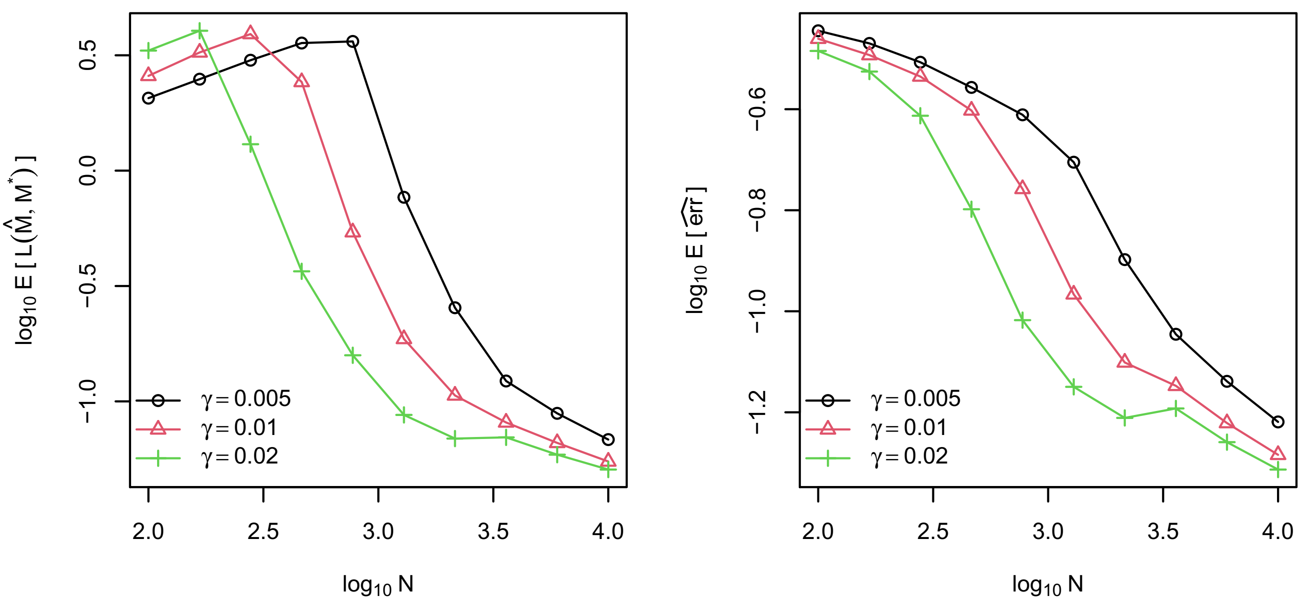


**Figure 2.** Average Kullback-Leibler divergence (left) and average link prediction error (right) of $\hat{M}$ for different choices of $N$ and $\gamma$.

### 4.2 Real data example: knowledge base completion

WordNet [1] is a large lexical knowledge base for English. It has been used in word sense disambiguation, text classification, question answering, and many other tasks in natural language processing [3, 5]. The basic components of WordNet are groups of words. Each group, called a synset, describes a distinct concept. In WordNet, synsets are linked by conceptual-semantic and lexical relations such as super-subordinate relation and antonym. We model WordNet as an MRN with the synsets as entities and the links between synsets as relations.

Following Bordes et al. [7], we use a subset of WordNet for analysis. The dataset contains 40,943 synsets and 18 types of relations. A triple $(i, j, k)$ is called valid if relation $k$ from entity $i$ to entity $j$ exists, i.e., $Y_{ijk} = 1$. All the other triples are called invalid triples. Among more than $3.0 \times 10^{10}$ possible triples in WordNet, only 151,442 triples are valid. We assume that 141,442 valid triples and the same proportion of invalid triples are observed. The goal of our analysis is to recover the unobserved part of the knowledge base. We adopt the ranking procedure, which is commonly used in knowledge graph embedding literature, to evaluate link predictions. Given a valid triple $\lambda = (i, j, k)$, we rank estimated scores for all the invalid triples inside $\Lambda_{\cdot jk} = \{(i', j, k) \mid i' \in [N]\}$ in descending order and call the rank of $\phi(\hat{\boldsymbol{x}}_\lambda)$ as the head rank of $\lambda$, denoted by $H_\lambda$. Similarly, we can define the tail rank $T_\lambda$ and the relation rank $R_\lambda$ by ranking $\phi(\hat{\boldsymbol{x}}_\lambda)$ among the estimated scores of invalid triples in $\Lambda_{ij\cdot}$ and $\Lambda_{i\cdot k}$, respectively. For a set $V$ of valid triples, the prediction performance can be evaluated by rank-based criteria, mean rank (MR), mean reciprocal rank (MRR), and hits at $q$ (Hits@q), which are defined as

$$\mathrm{MR_E} = \frac{1}{2|V|}\sum_{\lambda \in V} H_\lambda + T_\lambda, \;\; \mathrm{MR_R} = \frac{1}{|V|}\sum_{\lambda \in V} R_\lambda,$$

$$\mathrm{MRR_E} = \frac{1}{2|V|}\sum_{\lambda \in V} \frac{1}{H_\lambda} + \frac{1}{T_\lambda}, \;\; \mathrm{MRR_R} = \frac{1}{|V|}\sum_{\lambda \in V} \frac{1}{R_\lambda},$$

and

$$\mathrm{Hits_E@}q = \frac{1}{2|V|}\sum_{\lambda \in V} \mathbf{1}_{\{H_\lambda \le q\}} + \mathbf{1}_{\{T_\lambda \le q\}}, \;\; \mathrm{Hits_R@}q = \frac{1}{|V|}\sum_{\lambda \in V} \mathbf{1}_{\{R_\lambda \le q\}}.$$

The subscripts E and R represent the criteria for predicting entities and relations, respectively. Models with higher MRRs, Hits@$q$'s or lower MRs are more preferable. In addition, MRR is more robust to outliers than MR.

The three models described in (4), (5), and (36) are considered in our data analysis and we refer to them as Model 1, 2 and 3, respectively. For each model, the latent dimension $d$ takes value from $\{50, 100, 150, 200, 250\}$. Due to the high dimensionality of the parameter space, $L_2$ penalized MLE is used to obtain the estimated latent attributes $\hat{\boldsymbol{x}}$, with tuning parameters $\rho_1 = 0$ and $\rho_2$ chosen from $\{0, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$ in (22). Since information criteria based dimension and tuning parameter selection is computationally intensive for dataset of this scale, we set aside 5,000 of the unobserved valid triples as a validation set and select the $d$ and $\rho_2$ that produce the smallest $\mathrm{MRR_E}$ on this validation set. The model with the selected $d$ and $\rho_2$ is then evaluated on the test set consisting of the rest 5,000 unobserved valid triples.

The computed evaluation criteria on the test set are listed in Table 1. The table also includes the selected $d$ and $\rho_2$ for each of the three score models. Models 2 and 3 generate similar performance. The MRRs for the two models are very close to 1, and the Hits@$q$'s are higher than 90%, suggesting that the two models can identify the valid triples very well. Although Model 1 is inferior to the other two models in terms of most of the criteria, it outperforms them in $\mathrm{MR_E}$. The results imply that Model 2 and Model 3 could perform extremely bad for a few triples.

In addition to Models 1–3, we also display the performance of the Canonical Polyadic (CP) decomposition [22] and a tensor factorization approach, RESCAL [23]. Their $\mathrm{MRR_E}$ and $\mathrm{Hits_E@}10$ results on the WordNet dataset are extracted from [12] and [13], respectively. Both methods, especially CP, are outperformed by Model 3.

**Table 1.** Results for WordNet data analysis. The results for CP and RESCAL are extracted from [12] and [13].

| Method | $(d, \rho_2)$ | $\text{MR}_\text{E}$ | $\text{MRR}_\text{E}$ | $\text{Hits}_\text{E}@10$ | $\text{MR}_\text{R}$ | $\text{MRR}_\text{R}$ | $\text{Hits}_\text{R}@1$ |
|---|---|---|---|---|---|---|---|
| Model 1 | $(100, 10^{-5})$ | 385 | 0.64 | 0.888 | 1.41 | 0.896 | 0.817 |
| Model 2 | $(250, 10^{-4})$ | 769 | 0.94 | 0.945 | 1.31 | 0.968 | 0.959 |
| Model 3 | $(200, 10^{-4})$ | 499 | 0.94 | 0.947 | 1.13 | 0.978 | 0.967 |
| CP | - | - | 0.075 | 0.125 | - | - | - |
| RESCAL | - | - | 0.890 | 0.928 | - | - | - |

## 5 CONCLUDING REMARKS

In this article, we focused on the recovery of large-scale MRNs with a small portion of observations. We studied a generalized latent space model where entities and relations are associated with latent attribute vectors and conducted statistical analysis on the error of recovery. MLEs and pMLEs over a compact space are considered to estimate the latent attributes and the edge probabilities. We established non-asymptotic upper bounds for estimation error in terms of tail probability and risk, based on which we then studied the asymptotic properties when the size of MRN and latent dimension go to infinity simultaneously. A matching lower bound up to a log factor is also provided.

We kept $\phi$ generic for theoretical development. The choice of $\phi$ is usually problem-specific in practice. How to develop a data-driven method for selecting an appropriate $\phi$ is an interesting problem to investigate in future works.

Besides the latent space models, sparsity [24] or clustering assumptions [25] have been used to impose low-dimensional structures in single-relational networks. An MRN can be seen as a combination of several heterogeneous single-relational networks. The distribution of edges may vary dramatically across relations. Therefore, it is challenging to impose appropriate sparsity or cluster structures on MRNs. More empirical and theoretical studies are needed to quantify the impact of heterogeneous relations and to incorporate the information for recovering MRNs.

## APPENDIX

PROOF OF LEMMA 1. Let $\Theta_t = \{\boldsymbol{x} \in \Theta : L(M(\boldsymbol{x}), M^*) \geq t\}$ and $f(\boldsymbol{x}) = l(\boldsymbol{x}; Y_\mathcal{S}) - l(\boldsymbol{x}^*; Y_\mathcal{S})$ be the log likelihood ratio. Therefore, $f$ is a random field living on $\Theta$. By writing $f(\boldsymbol{x})$, we omit the second argument. In explicit form, $f(\boldsymbol{x}) = \sum_{\lambda \in \Lambda} Z_\lambda$, where

$$Z_\lambda = 1_{\lambda \in \mathcal{S}} \left[ Y_\lambda \log \frac{M_\lambda(\boldsymbol{x})}{M^*_\lambda} + (1 - Y_\lambda) \log \frac{1 - M_\lambda(\boldsymbol{x})}{1 - M^*_\lambda} \right]. \tag{37}$$

We have $E[Z_\lambda] = -\gamma D\left(M^*_\lambda || M_\lambda(\boldsymbol{x})\right)$ and $|Z_\lambda| \leq C$. It follows that $f$ has properties (i) $f(\boldsymbol{x}^*) = 0$, (ii) $f(\hat{\boldsymbol{x}}) \geq 0$, (iii) $E[f(\boldsymbol{x})] = -nL(M(\boldsymbol{x}), M^*)$. Based on the definition of $\Theta_t$ and property (ii), we have

$$P\left(L(\hat{M}, M^*) \geq t\right) = P(\hat{\boldsymbol{x}} \in \Theta_t) \leq P\left(\sup_{\boldsymbol{x} \in \Theta_t} f(\boldsymbol{x}) \geq 0\right). \tag{38}$$

From property (iii), we get that

$$E[f(\boldsymbol{x})] \leq -nt, \quad \forall \boldsymbol{x} \in \Theta_t. \tag{39}$$

According to Lemma 3 in Appendix, when $C \geq 2$, the variance of $Z_\lambda$ is bounded by

$$\mathrm{Var}\left[Z_\lambda\right] = \gamma M_\lambda^*(1-M_\lambda^*)\left(\log\frac{M_\lambda}{1-M_\lambda} - \log\frac{M_\lambda^*}{1-M_\lambda^*}\right)^2 \leq 2\gamma C D\left(M_\lambda^* || M_\lambda\right).$$

It follows that

$$\mathrm{Var}\left[f(\boldsymbol{x})\right] = \sum_{\lambda\in\Lambda}\mathrm{Var}\left[Z_\lambda\right] \leq 2\gamma C \sum_{\lambda\in\Lambda} D\left(M_\lambda^*||M_\lambda\right) = -2CE\left[f(\boldsymbol{x})\right]. \tag{40}$$

By Bennett's inequality,

$$P\left(f(\boldsymbol{x}) \geq -s\right) \leq \exp\left\{\frac{s+E\left[f(\boldsymbol{x})\right]}{C} h\left(-\frac{C\left[s+E\left[f(\boldsymbol{x})\right]\right]}{\mathrm{Var}\left[f(\boldsymbol{x})\right]}\right)\right\}, \tag{41}$$

where $0 < s < nt$ and $h(u) = \left(1+\frac{1}{u}\right)\log\left(1+u\right) - 1$ is an increasing function for $u > 0$. Hence by bounds in (39)(40),

$$P\left(f(\boldsymbol{x}) \geq -s\right) \leq \exp\left\{-\frac{nt-s}{C} h\left(\frac{s+E\left[f(\boldsymbol{x})\right]}{2E\left[f(\boldsymbol{x})\right]}\right)\right\} \leq \exp\left\{-\frac{nt-s}{C} h\left(\frac{1}{2} - \frac{s}{2nt}\right)\right\}. \tag{42}$$

Let $\boldsymbol{z} = \mathrm{argmax}_{\boldsymbol{x}\in\Theta_t} f(\boldsymbol{x})$ be the random vector on $\Theta_t$ where $f(\boldsymbol{x})$ reaches its maximum. Let $\mathcal{N}_{\epsilon,\mathcal{E}}$ and $\mathcal{N}_{\epsilon,\mathcal{R}}$ be the $\epsilon$-covering centers for $\mathcal{E}$ and $\mathcal{R}$ respectively. Since $\mathcal{E}$ and $\mathcal{R}$ are balls of radius $U$, we can find $\epsilon$-coverings such that $|\mathcal{N}_{\epsilon,\mathcal{E}}| \leq \left(1+2U/\epsilon\right)^{d_E}$ and $|\mathcal{N}_{\epsilon,\mathcal{R}}| \leq \left(1+2U/\epsilon\right)^{d_R}$. For $\boldsymbol{z} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N, \boldsymbol{w}_1, \ldots, \boldsymbol{w}_K)$, there exists some $\boldsymbol{x} = \left(\boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_N', \boldsymbol{w}_1', \ldots, \boldsymbol{w}_K'\right) \in \mathcal{N}_{\epsilon,\mathcal{E}}^N \times \mathcal{N}_{\epsilon,\mathcal{R}}^K$ such that $\|\boldsymbol{\theta}_i' - \boldsymbol{\theta}_i\| \leq \epsilon, \forall i \in [N]$ and $\|\boldsymbol{w}_k' - \boldsymbol{w}_k\| \leq \epsilon, \forall k \in [K]$. Therefore,

$$f(\boldsymbol{z}) - f(\boldsymbol{x}) \leq \sum_{\lambda\in\mathcal{S}} |\phi(\boldsymbol{z}_\lambda) - \phi(\boldsymbol{x}_\lambda)| \leq \alpha \sum_{\lambda\in\mathcal{S}} \|\boldsymbol{z}_\lambda - \boldsymbol{x}_\lambda\| \leq \sqrt{3}\alpha|\mathcal{S}|\epsilon. \tag{43}$$

By Bennett's inequality, for every $\beta > 0$,

$$p\left(|\mathcal{S}| - n > n\beta\right) \leq \exp\left\{-n\beta h\left(\frac{\beta}{1-\gamma}\right)\right\} \leq \exp\left\{-n\beta h(\beta)\right\}. \tag{44}$$

When $|\mathcal{S}| \leq n(1+\beta)$, set $\epsilon = \frac{s}{\sqrt{3}\alpha n(1+\beta)}$, then $f(\boldsymbol{z}) - f(\boldsymbol{x}) \leq s$. Combining (38) (42) and (44), we get that

$$\begin{aligned}
P\left(L(\hat{M}, M^*) \geq t\right) &\leq P\left(\sup_{\boldsymbol{x}\in\Theta_t} f(\boldsymbol{x}) \geq 0, |\mathcal{S}| \leq n(1+\beta)\right) + P\left(|\mathcal{S}| > n(1+\beta)\right) \\
&\leq P\left(\max_{\boldsymbol{x}\in\mathcal{N}_{\epsilon,\mathcal{E}}^N\times\mathcal{N}_{\epsilon,\mathcal{R}}^K} f(\boldsymbol{x}) \geq -s, |\mathcal{S}| \leq n(1+\beta)\right) + P\left(|\mathcal{S}| > n(1+\beta)\right) \\
&\leq |\mathcal{N}_{\epsilon,\mathcal{E}}^N \times \mathcal{N}_{\epsilon,\mathcal{R}}^K| \max_{\boldsymbol{x}\in\mathcal{N}_{\epsilon,\mathcal{E}}^N\times\mathcal{N}_{\epsilon,\mathcal{R}}^K} P\left(f(\boldsymbol{x}) \geq -s\right) + \exp\left\{-n\beta h(\beta)\right\} \\
&\leq \exp\left\{-\frac{nt-s}{C} h\left(\frac{1}{2} - \frac{s}{2nt}\right)\right\}\left(1 + \frac{2\sqrt{3}\alpha U n(1+\beta)}{s}\right)^m + \exp\left\{-n\beta h(\beta)\right\},
\end{aligned} \tag{45}$$

where $m = Nd_E + Kd_R$ is the degree of freedom.

PROOF OF LEMMA 2. To bound $E\left[L(\hat{M}, M^*)\right]$, set $s = \frac{1}{2}nt$ and $\beta = 1 + t$ in (14) to get

$$P\left(L(\hat{M}, M^*) \geq t\right) \leq \exp\left\{-\frac{nt}{C_1}\right\}\left(1 + \frac{C_2}{2} + \frac{C_2}{t}\right)^m + \exp\left\{-\frac{1}{3}n(1+t)\right\}. \tag{46}$$

By Fubini's Theorem,

$$E\left[L(\hat{M}, M^*)\right] = \int_0^\infty P\left(L(\hat{M}, M^*) \geq t\right) dt \leq t_0 + \int_{t_0}^\infty P\left(L(\hat{M}, M^*) \geq t\right) dt. \tag{47}$$

Let $C_3 = 2\max\left[\{C_1, C_2\}\right]$ and $t_0 = C_3 \frac{m}{n} \log \frac{n}{m}$. When $t \geq t_0$ and $\frac{n}{m} \geq C_2 + e$,

$$1 + \frac{C_2}{2} + \frac{C_2}{t} \leq 1 + \frac{C_2}{2} + \frac{C_2 n}{C_3 m \log \frac{n}{m}} \leq 1 + \frac{C_2}{2} + \frac{n}{2m} \leq \frac{n}{m}. \tag{48}$$

Thus

$$P\left(L(\hat{M}, M^*) \geq t\right) \leq \exp\left\{-\frac{nt}{C_1} + m \log \frac{n}{m}\right\} + \exp\left\{-\frac{1}{3}n(1+t)\right\}, \quad t \geq t_0. \tag{49}$$

Hence by (47) and (49),

$$\begin{aligned} E\left[L(\hat{M}, M^*)\right] &\leq t_0 + \frac{C_1}{n} \exp\left\{-\frac{nt_0}{C_1} + m \log \frac{n}{m}\right\} + \frac{3}{n} \exp\left\{-\frac{1}{3}n(1+t_0)\right\} \\ &\leq C_3 \frac{m}{n} \log \frac{n}{m} + \frac{C_1}{n} \exp\left\{-m \log \frac{n}{m}\right\} + \frac{3}{n} \exp\left\{-\frac{1}{3}\left(n + C_3 m \log \frac{n}{m}\right)\right\}. \end{aligned} \tag{50}$$

PROOF OF THEOREM 1. When $t$ is a constant, let $s$ be absolute constant and $\beta = m \to \infty$ in Lemma 1. We analyze the order of three exponential terms on the right side of (14),

$$-\frac{nt - s}{C} h\left(\frac{1}{2} - \frac{s}{2nt}\right) \sim -\frac{h\left(\frac{1}{2}\right)}{C} nt,$$

$$m \log\left(1 + \frac{2\sqrt{3}\alpha U n(1+\beta)}{s}\right) \sim m \log(mn),$$

$$-n\beta h(\beta) \sim -nm \log m.$$

Hence, both the second and the third term is asymptotically ignorable compared to the first term. It follows that

$$\log P\left(L(\hat{M}, M^*) \geq t\right) \lesssim -\frac{h\left(\frac{1}{2}\right)}{C} nt.$$

When $t = \frac{2C}{h(\frac{1}{2})}\frac{m}{n}\log\frac{n}{m}$, let $s = m$ and $\beta$ be absolute constant. The exponential terms

$$-\frac{nt-s}{C}h\left(\frac{1}{2}-\frac{s}{2nt}\right) \sim -2m\log\frac{n}{m},$$
$$m\log\left(1+\frac{2\sqrt{3}\alpha Un(1+\beta)}{s}\right) = m\log\frac{n}{m}+O(m).$$

The third term $\exp\{-n\beta h(\beta)\}$ is negligible. Therefore,

$$\log P\left(L(\hat{M},M^*)\geq t\right) \lesssim -m\log\frac{n}{m}. \tag{51}$$

To bound the risk, we use similar approach as proof of Lemma 2. Let $s = m$, $\beta = 1+t$ and $t_0 = \frac{2C}{h(\frac{1}{2})}\frac{m}{n}\log\frac{n}{m}$.

$$\int_{t_0}^{\infty}\exp\left\{-\frac{nt-s}{C}h\left(\frac{1}{2}-\frac{s}{2nt}\right)\right\}dt \leq \frac{C}{nh\left(\frac{1}{2}-\frac{s}{2nt_0}\right)}\exp\left\{-\frac{nt_0-s}{C}h\left(\frac{1}{2}-\frac{s}{2nt_0}\right)\right\}$$
$$\sim \frac{C}{nh\left(\frac{1}{2}\right)}\exp\left\{-2m\log\frac{n}{m}\right\},$$

$$m\log\left(1+\frac{2\sqrt{3}\alpha Un(1+\beta)}{s}\right) \leq m\log\left(1+\frac{2\sqrt{3}\alpha Un(2+t_0)}{m}\right) \sim m\log\frac{n}{m},$$

and

$$\int_{t_0}^{\infty}\exp\left\{-n(1+t)h\left(1+t\right)\right\}dt \leq \frac{3}{n}\exp\left\{-\frac{1}{3}n(1+t_0)\right\} = o\left(\exp\left\{-m\log\frac{n}{m}\right\}\right).$$

It follows that

$$\begin{aligned} E\left[L(\hat{M},M^*)\right] &\leq t_0+\int_{t_0}^{\infty}P\left(L(\hat{M},M^*)\geq t\right)dt \\ &\lesssim t_0+o\left(t_0\right) \sim \frac{2C}{h\left(\frac{1}{2}\right)}\frac{m}{n}\log\frac{n}{m}. \end{aligned} \tag{52}$$

Since $h(\frac{1}{2}) \geq \frac{1}{5}$, we proof the results.

LEMMA 3. *$\forall x, y \in [-C, C]$, we have*

$$\sigma(x)\left(1-\sigma(x)\right)\left(y-x\right)^2 \leq 2\max\left\{C,2\right\}D\left(\sigma(x)||\sigma(y)\right), \tag{53}$$

PROOF. We only need to show the result for $x \geq 0$ by symmetry. For any fixed $x \in [0, C]$, define $g(y) = 2C_mD\left(\sigma(x)||\sigma(y)\right) - \sigma(x)\left(1-\sigma(x)\right)\left(y-x\right)^2$, where $C_m = \max\{C, 2\}$. Since

$$g'(y) = 2C_m(\sigma(y)-\sigma(x)) - 2\sigma(x)(1-\sigma(x))(y-x), \tag{54}$$

we have $g'(x) = g(x) = 0$. It remains to show that $\frac{g'(y)}{y-x} > 0$ for all $y \in [-C, C] \setminus \{x\}$, then $g(x)$ reaches the minimum at $x = 0$ and $g(y) \geq 0$ on $[-C, C]$. Equivalently, we want to show that

$$C_m(\sigma(y) - \sigma(x))/(y - x) > \sigma(x)(1 - \sigma(x)).$$

Note that $(\sigma(y) - \sigma(x))/(y - x)$ is the slope of secant line on logistic function and reaches its minimum at $y = C$. It suffices to show that

$$(C - x)\sigma(x)(1 - \sigma(x)) + C_m\sigma(x) \leq C_m\sigma(C), \forall x \in [0, C] \tag{55}$$

Let $h(x)$ be left side above. By taking the derivative, we get

$$h'(x) = [C_m - 1 - (C - x)(2\sigma(x) - 1)]\, \sigma(x)\, (1 - \sigma(x)).$$

If $1 \leq x \leq C$, then $(C - x)(2\sigma(x) - 1) \leq C - 1 \leq C_m - 1$. If $0 \leq x \leq 1$, then $(C - x)(2\sigma(x) - 1) \leq C(2\sigma(1) - 1) \leq \frac{1}{2}C \leq C_m - 1$. Therefore, $h'(x) \geq 0$ on $[0, C]$. It follows that $h(x) \leq h(C) = C_m\sigma(C)$.

To prove the lower bound in Theorem 2, we will use Lemma 4 – 6. Since Lemma 4 [26] and Lemma 5 [27] are well established results in literature, we will skip the proofs.

LEMMA 4 (Gilbert-Varshamov bound). *There exists a subset $\mathcal{V}$ of the $d$-dimensional hypercube $\{-1, 1\}^d$ of size at least $\exp\{d/8\}$ such that the Hamming distance*

$$\sum_{i=1}^{d} 1_{\boldsymbol{u}_i \neq \boldsymbol{v}_i} \geq \frac{1}{4}d \tag{56}$$

*for all $\boldsymbol{u} \neq \boldsymbol{v}$ with $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{V}$.*

LEMMA 5 (Fano's inequality). *Let $V$ be a uniform random variable taking values in a finite set $\mathcal{V}$ with cardinality $|\mathcal{V}| \geq 2$. For any Markov chain $V \to X \to \hat{V}$,*

$$P\left(\hat{V} \neq V\right) \geq 1 - \frac{I(V; X) + \log 2}{\log(|\mathcal{V}|)}, \tag{57}$$

*where $I(V; X)$ is the mutual information between $V$ and $X$.*

LEMMA 6. *Suppose that $p, q \in (0, 1)$. Then*

$$D(p||q) \leq \frac{(p - q)^2}{q(1 - q)}. \tag{58}$$

PROOF. Since $D(1-p||1-q) = D(p||q)$, it suffices to show for case $p \leq q$. View $D(p||q)$ as a function of $q$. By mean value theorem, there exists $\xi \in [p, q]$ such that

$$D(p||q) - D(p||p) = \frac{\xi - p}{\xi(1 - \xi)}(q - p) \tag{59}$$

Note that $\frac{\xi-p}{\xi(1-\xi)}$ is increasing in $\xi$ and $D(p||p) = 0$. Hence, $D(p||q) \leq \frac{(q-p)^2}{q(1-q)}$.

PROOF OF THEOREM 2. Let $\boldsymbol{u}_0 = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_0', \boldsymbol{w}_0)$, $\tilde{\boldsymbol{x}} = (\underbrace{\boldsymbol{\theta}_0, \ldots, \boldsymbol{\theta}_0}_{\lfloor \frac{N}{2} \rfloor}, \underbrace{\boldsymbol{\theta}_0', \ldots, \boldsymbol{\theta}_0'}_{\lceil \frac{N}{2} \rceil}, \underbrace{\boldsymbol{w}_0, \ldots, \boldsymbol{w}_0}_{K})$ and

$$\tilde{\Lambda} = \left\{ (i,j,k) \in \Lambda \mid i \leq \lfloor \frac{N}{2} \rfloor, j > \lfloor \frac{N}{2} \rfloor \right\} \subset \Lambda$$

with cardinality $|\tilde{\Lambda}| = \lfloor \frac{N}{2} \rfloor \lceil \frac{N}{2} \rceil K$. If $\boldsymbol{x} \in \mathcal{N}_r(\tilde{\boldsymbol{x}})$, then $\boldsymbol{x}_\lambda \in \mathcal{N}_r(\boldsymbol{u}_0)$ for every $\lambda \in \tilde{\Lambda}$. Hence according to Assumption 3,

$$\left| \sigma\left(\phi(\boldsymbol{x}_\lambda)\right) - \sigma\left(\phi(\boldsymbol{x}_\lambda')\right) \right| \geq \kappa \|\boldsymbol{x}_\lambda - \boldsymbol{x}_\lambda'\|, \quad \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{N}_r(\tilde{\boldsymbol{x}}), \lambda \in \tilde{\Lambda}. \tag{60}$$

We will find $\boldsymbol{x}^*$ in the vicinity of $\tilde{\boldsymbol{x}}$ such that (20) holds.

Let $\mathcal{H}_E = \left\{-\delta/\sqrt{d_E}, \delta/\sqrt{d_E}\right\}^{Nd_E}$ and $\mathcal{H}_R = \left\{-\delta/\sqrt{d_R}, \delta/\sqrt{d_R}\right\}^{Kd_R}$ be two hypercubes. According to Gilbert-Varshamov bound in Lemma 4, there exist $\mathcal{V}_E \subset \mathcal{H}_E$ and $\mathcal{V}_R \subset \mathcal{H}_R$ such that $|\mathcal{V}_E| \geq \exp\{Nd_E/8\}$, $|\mathcal{V}_R| \geq \exp\{Kd_R/8\}$ and

$$\sum_{i=1}^{Nd_E} 1_{\boldsymbol{u}_i \neq \boldsymbol{v}_i} \geq \frac{1}{4} N d_E, \quad \forall \boldsymbol{u}, \boldsymbol{v} \in \mathcal{V}_E, \boldsymbol{u} \neq \boldsymbol{v}, \tag{61}$$

$$\sum_{i=1}^{Kd_R} 1_{\boldsymbol{u}_i \neq \boldsymbol{v}_i} \geq \frac{1}{4} K d_R, \quad \forall \boldsymbol{u}, \boldsymbol{v} \in \mathcal{V}_R, \boldsymbol{u} \neq \boldsymbol{v}. \tag{62}$$

For $\boldsymbol{u} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N) \in \mathcal{V}_E$, $\boldsymbol{v} = (\boldsymbol{\theta}_1', \ldots, \boldsymbol{\theta}_N') \in \mathcal{V}_E$ and $\boldsymbol{u} \neq \boldsymbol{v}$, (61) suggests that

$$\sum_{i=1}^{N} \|\boldsymbol{\theta}_i - \boldsymbol{\theta}_i'\|^2 \geq \sum_{i=1}^{N} \left(2\delta/\sqrt{d_E}\right)^2 \frac{1}{4} N d_E = N\delta^2, \tag{63}$$

Likewise, from (62) we can get that

$$\sum_{i=1}^{K} \|\boldsymbol{w}_k - \boldsymbol{w}_k'\| \geq K\delta^2, \tag{64}$$

with $\boldsymbol{u} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K) \in \mathcal{V}_R$, $\boldsymbol{v} = (\boldsymbol{w}_1', \ldots, \boldsymbol{w}_K') \in \mathcal{V}_R$ and $\boldsymbol{u} \neq \boldsymbol{v}$.

Let $\mathcal{V} = \{\tilde{\boldsymbol{x}} + \boldsymbol{e} \mid \boldsymbol{e} \in \mathcal{V}_E \times \mathcal{V}_R\} = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(T)}\}$ where $T = |\mathcal{V}_E||\mathcal{V}_R| \geq \exp\{m/8\}$. By the definition of $\delta$-neighborhood and size of hypercubes, we have $\mathcal{V} \subset \mathcal{N}_\delta(\tilde{\boldsymbol{x}})$ and thus property in (60) holds for $\delta \leq r$. The corresponding tensors are denoted as $M(\mathcal{V}) = \left\{M^{(1)}, \ldots, M^{(T)}\right\}$ where $M^{(i)} = M\left(\boldsymbol{x}^{(i)}\right)$ for $i \in [T]$. Let $\boldsymbol{z} = \operatorname{argmin}_{\boldsymbol{x} \in \mathcal{V}} \|\hat{M} - M(\boldsymbol{x})\|$, thus $M(\boldsymbol{z})$ is the closet tensor to $\hat{M}$ in $M(\mathcal{V})$ under Frobenius norm. By triangular inequality,

$$\|\hat{M} - M^{(i)}\| \geq \frac{1}{2}\left(\|\hat{M} - M^{(i)}\| + \|\hat{M} - M(\boldsymbol{z}\|)\right) \geq \frac{1}{2}\|M^{(i)} - M(\boldsymbol{z})\|, \quad \forall i \in [T]. \tag{65}$$

Note that $\boldsymbol{z}, \boldsymbol{x}^{(i)} \in \mathcal{V}$, according to Pinsker's inequality and (60),

$$L\left(\hat{M}, M^{(i)}\right) \geq \frac{2}{|\Lambda|}\|\hat{M} - M^{(i)}\|^2 \geq \frac{1}{2|\Lambda|}\|M^{(i)} - M(\boldsymbol{z})\|^2 \geq \frac{\kappa^2}{2|\Lambda|}\sum_{\lambda\in\tilde{\Lambda}}\|\boldsymbol{x}_\lambda^{(i)} - \boldsymbol{z}_\lambda\|^2.$$

For all $\boldsymbol{x} \neq \boldsymbol{x}'$ with $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{V}$ and $N \geq 2$,

$$\begin{aligned}\frac{1}{|\Lambda|}\sum_{\lambda\in\tilde{\Lambda}}\|\boldsymbol{x}_\lambda - \boldsymbol{x}'_\lambda\|^2 &\geq \frac{1}{|\Lambda|}\left(\lfloor\frac{N}{2}\rfloor K\sum_{i\in[N]}\|\boldsymbol{\theta}_i - \boldsymbol{\theta}'_i\|^2 + \lfloor\frac{N}{2}\rfloor\lceil\frac{N}{2}\rceil\sum_{k\in[K]}\|\boldsymbol{w}_k - \boldsymbol{w}'_k\|^2\right)\\ &\geq \min\left\{\frac{1}{3}\frac{1}{N}\sum_{i\in[N]}\|\boldsymbol{\theta}_i - \boldsymbol{\theta}'_i\|^2, \frac{2}{9}\frac{1}{K}\sum_{k\in[K]}\|\boldsymbol{w}_k - \boldsymbol{w}'_k\|^2\right\} = \frac{2}{9}\delta^2.\end{aligned} \tag{66}$$

Hence when $\boldsymbol{x}^{(i)} \neq \boldsymbol{z}$,

$$L\left(\hat{M}, M^{(i)}\right) \geq \frac{1}{9}\kappa^2\delta^2. \tag{67}$$

Let $P_i$ denote the probability measure under $\boldsymbol{x}^{(i)}$. Results above show that

$$P_i\left(L(\hat{M}, M^{(i)}) \geq \frac{1}{9}\kappa^2\delta^2\right) \geq P_i\left(\boldsymbol{x}^{(i)} \neq \boldsymbol{z}\right), \quad \forall i \in [N]. \tag{68}$$

Assign a prior on $\boldsymbol{x}$ that is uniform on $\mathcal{V}$ and denote by $P_{\mathcal{V}}$ the Bayes average probability with respect to the prior. By Fano's inequality in Lemma 5,

$$P_{\mathcal{V}}\left(\boldsymbol{z} \neq \boldsymbol{x}\right) \geq 1 - \frac{I(\boldsymbol{x}; Y_{\mathcal{S}}) + \log 2}{\log|T|}, \tag{69}$$

where $I(\boldsymbol{x}; X_{\mathcal{S}})$ is the mutual information between $\boldsymbol{x}$ and $Y_{\mathcal{S}}$. It can be bounded by the maximum pairwise KL divergence of $Y_{\mathcal{S}}$ under $P_i$ and $P_j$ as follows,

$$\begin{aligned}I(\boldsymbol{x}, Y_{\mathcal{S}}) =& \frac{1}{T}\sum_{i=1}^{T} D\left(P_i(Y_{\mathcal{S}})||P_{\mathcal{V}}(Y_{\mathcal{S}})\right) \leq \max_{i\neq j} D\left(P_i(Y_{\mathcal{S}})||P_j(Y_{\mathcal{S}})\right) =\\ &\max_{i\neq j}\sum_{\lambda\in\Lambda} D\left(P_i(Y_\lambda, \lambda\in\mathcal{S})||P_j(Y_\lambda, \lambda\in\mathcal{S})\right) = \max_{i\neq j} nL\left(M^{(i)}, M^{(j)}\right).\end{aligned} \tag{70}$$

Since $\sigma(\cdot)$ is logistic function, the derivative $\sigma'(x) = \sigma(x)\left(1 - \sigma(x)\right) < 1$. By Assumption 2, $\phi(\cdot)$ is Lipschitz continuous with coefficient $\alpha$ , we get that $\sigma(\phi(\cdot))$ is also Lipschitz continuous with coefficient $\alpha$. Let $b = \sup_{\boldsymbol{u}\in\mathcal{N}_r(\boldsymbol{u}_0)} \sigma\left(\phi(\boldsymbol{u})\right)$, by Lemma 6 we get

$$L(M^{(i)}, M^{(j)}) \leq \frac{\|M^{(i)} - M^{(j)}\|^2}{|\Lambda|b(1-b)} \leq \frac{\alpha^2\sum_{\lambda\in\Lambda}\|\boldsymbol{x}_\lambda^{(i)} - \boldsymbol{x}_\lambda^{(j)}\|^2}{|\Lambda|b(1-b)} \leq \frac{3(2\delta)^2\alpha^2}{b(1-b)} = \frac{12\alpha^2\delta^2}{b(1-b)} \tag{71}$$

for all $i, j \in [N]$. Hence, there exists $\boldsymbol{x}^{(i)} \in \mathcal{V}$ such that

$$P_i\left(\boldsymbol{z} \neq \boldsymbol{x}^{(i)}\right) \geq 1 - \frac{\frac{12\alpha^2\delta^2 n}{b(1-b)} + \log 2}{\log |T|} \geq 1 - \frac{\frac{12\alpha^2\delta^2 n}{b(1-b)} + 1}{m/8}. \tag{72}$$

Let $\boldsymbol{x}^* = \boldsymbol{x}^{(i)}$ , $P = P_i$ and

$$\delta^2 = \frac{(m/16-1)b(1-b)}{12\alpha^2 n} \leq r^2.$$

It follows from (68) that

$$P\left(L(\hat{M}, M^{(i)}) \geq \frac{\kappa^2 b(1-b)}{108\alpha^2}\frac{m/16-1}{n}\right) \geq \frac{1}{2}. \tag{73}$$

PROOF OF THEOREM 3. We will show the result by continuing the proof of Lemma 1 and Theorem 1 with some modifications. Let $f_\rho(\boldsymbol{x})$ be the penalized log likelihood ratio, we have

$$\begin{aligned} f_\rho(\boldsymbol{x}) &= l_\rho\left(\boldsymbol{x}; Y_{\mathcal{S}}\right) - l_\rho\left(\boldsymbol{x}^*; Y_{\mathcal{S}}\right) \\ &= f(\boldsymbol{x}) - \rho_1\left(\|\boldsymbol{x}\|_1 - \|\boldsymbol{x}^*\|_1\right) - \rho_2\left(\|\boldsymbol{x}\|^2 - \|\boldsymbol{x}^*\|^2\right) \\ &\leq f(\boldsymbol{x}) + \sqrt{2}\rho_1(N+K)U + \rho_2(N+K)U^2 \end{aligned} \tag{74}$$

According to (43), there exists $\boldsymbol{x}$ among the $\epsilon$-covering centers such that

$$\begin{aligned} f_\rho(\boldsymbol{z}) - f_\rho(\boldsymbol{x}) &= f(\boldsymbol{z}) - f(\boldsymbol{x}) - \rho_1\left(\|\boldsymbol{z}\|_1 - \|\boldsymbol{x}\|_1\right) - \rho_2\left(\|\boldsymbol{z}\|^2 - \|\boldsymbol{x}\|^2\right) \\ &\leq \sqrt{3}\alpha|\mathcal{S}|\epsilon + \sqrt{2}\rho_1(N+K)\epsilon + 2\rho_2(N+K)U\epsilon, \end{aligned} \tag{75}$$

where $\boldsymbol{z} = \operatorname{argmax}_{\boldsymbol{x}\in\Theta_t} f_\rho(\boldsymbol{x})$. It follow that when $|\mathcal{S}| \leq n(1+\beta)$ and $f_\rho(\boldsymbol{z}) \geq 0$,

$$\begin{aligned} f_\rho(\boldsymbol{x}) &\geq -\sqrt{3}\alpha|\mathcal{S}|\epsilon - \sqrt{2}\rho_1(N+K)\epsilon - 2\rho_2(N+K)U\epsilon \\ &\geq -s - \frac{(N+K)s}{\alpha n(1+\beta)}\left(\sqrt{\frac{2}{3}}\rho_1 + \frac{2}{\sqrt{3}}\rho_2 U\right), \end{aligned} \tag{76}$$

with $\epsilon = \frac{s}{\sqrt{3}\alpha n(1+\beta)}$. Hence, we can rewrite (45) as

$$\begin{aligned} P\left(L(\hat{M}, M^*) \geq t\right) &\leq P\left(\sup_{\boldsymbol{x}\in\Theta_t} f_\rho(\boldsymbol{x}) \geq 0, |\mathcal{S}| \leq n(1+\beta)\right) + P\left(|\mathcal{S}| > n(1+\beta)\right) \\ &\leq |\mathcal{N}_{\epsilon,\mathcal{E}}^N \times \mathcal{N}_{\epsilon,\mathcal{R}}^K| P\left(f(\boldsymbol{x}) \geq -s_\rho\right) + \exp\left\{-n\beta h(\beta)\right\} \\ &\leq \exp\left\{-\frac{nt - s_\rho}{C} h\left(\frac{1}{2} - \frac{s_\rho}{2nt}\right)\right\}\left(1 + \frac{2\sqrt{3}\alpha U n(1+\beta)}{s}\right)^m + \exp\left\{-n\beta h(\beta)\right\}, \end{aligned} \tag{77}$$

where

$$s_\rho = s + \frac{(N+K)s}{\alpha n(1+\beta)}\left(\sqrt{\frac{2}{3}}\rho_1 + \frac{2}{\sqrt{3}}\rho_2 U\right) + \sqrt{2}\rho_1(N+K)U + \rho_2(N+K)U^2.$$

Therefore, $s_\rho = s + o(s) + O(N) = o(nt)$ when $t$ and $s$ are absolute constant or when $t = \frac{2C}{h(\frac{1}{2})}\frac{m}{n}\log\frac{n}{m}$ and $s = m$. Hence the proof of Theorem 1 applies and the asymptotic results hold.

PROOF OF COROLLARY 1, 2 AND 3. To show these corollaries, we associate $MSE_\phi$ and $\widehat{err}$ with $L(\hat{M}, M^*)$. The first and second order derivatives of $D\left(\sigma(x)||\sigma(y)\right)$ as a function of $y$ are

$$\frac{\partial}{\partial y}D\left(\sigma(x)||\sigma(y)\right) = \sigma(y) - \sigma(x), \quad \frac{\partial^2}{\partial^2 y}D\left(\sigma(x)||\sigma(y)\right) = \sigma(y)\left(1 - \sigma(y)\right). \tag{78}$$

By Taylor expansion, there exists $\xi = ux + (1-u)y$ with $u \in (0,1)$ such that $D\left(\sigma(x)||\sigma(y)\right) = \frac{1}{2}\sigma(\xi)\left(1-\sigma(\xi)\right)(y-x)^2$. Hence, for $x, y \in [-C, C]$,

$$\frac{1}{2}\sigma(C)\left(1-\sigma(C)\right)(y-x)^2 \leq D\left(\sigma(x)||\sigma(y)\right) \leq \frac{1}{8}(y-x)^2. \tag{79}$$

It follows that

$$\frac{1}{2}\sigma(C)\left(1-\sigma(C)\right)MSE_\phi \leq L\left(\hat{M}, M^*\right) \leq \frac{1}{8}MSE_\phi. \tag{80}$$

where $MSE_\phi = \frac{1}{|\Lambda|}\sum_{\lambda\in\Lambda}\left(\phi(\hat{\boldsymbol{x}}_\lambda) - \phi(\boldsymbol{x}^*_\lambda)\right)^2$ is the mean squared error of edge scores. The upper bound of $MSE_\phi$ follows from Theorem 3 and left half of (80). By Theorem 2 and right half of (80), we get the corresponding lower bound. Likewise, for $\widehat{err}$ we can derive the upper bound by

$$L\left(\hat{M}, M^*\right) = \frac{1}{|\Lambda|}\sum_{\lambda\in\Lambda}D\left(M^*_\lambda||\hat{M}_\lambda\right) \geq \frac{1}{|\Lambda|}\sum_{\lambda\in\Lambda}1_{\hat{Y}_\lambda\neq Y^*_\lambda}D\left(\frac{1}{2}+\varepsilon||\frac{1}{2}\right) \geq 2\epsilon^2\widehat{err}. \tag{81}$$

PROOF OF THEOREM 4. Let $\Theta_\tau = \left\{\boldsymbol{x} \in \mathcal{E}^N \times \mathcal{R}^K \mid \|\boldsymbol{x}\|_0 \leq m_\tau\right\}$ be subspaces of $\Theta$ with at most $m_\tau$ non-zeros and $\mathcal{N}_{\Theta_\tau}$ be its $\epsilon$-covering centers. There are $\binom{m}{m_\tau}$ combinations of support, and each subspace has a covering number of $\left(1+\frac{2U}{\epsilon}\right)^{m_\tau}$. Hence, the overall $\epsilon$-covering number of $\Theta_\tau$ would be

$$|\mathcal{N}_{\Theta_\tau}| = \binom{m}{m_\tau}\left(1+\frac{2U}{\epsilon}\right)^{m_\tau}. \tag{82}$$

We can rewrite Lemma 1 as

$$P\left(L(\hat{M}, M^*) \geq t\right) \leq \exp\left\{-\text{I} + \text{II}\right\} + \exp\left\{-\text{III}\right\}, \tag{83}$$

where

$$\text{I} = \frac{nt - s}{C}h\left(\frac{1}{2} - \frac{s}{2nt}\right),$$
$$\text{II} = \log\binom{m}{m_\tau} + m_\tau\log\left(1 + \frac{2\sqrt{3}\alpha Un(1+\beta)}{s}\right),$$
$$\text{III} = n\beta h(\beta).$$

By Stirling's approximation,

$$\begin{aligned}\log\binom{m}{m_\tau} &\sim -m_\tau\log\tau-(m-m_\tau)\log(1-\tau)-\frac{1}{2}\log m\\ &\lesssim m_\tau(-\log\tau+1)-\frac{1}{2}\log m=O(m_\tau).\end{aligned} \tag{84}$$

To get the results, when $t$ is absolute constant, let $s$ be absolute constant and $\beta = m$. When $t = \frac{2C}{h(\frac{1}{2})}\frac{m_\tau}{n}\log\frac{n}{m_\tau}$, let $s = m_\tau$ and $\beta$ be absolute constant. For risk upper bound, select $s = m_\tau, \beta = 1 + t$ and $t_0 = \frac{2C}{h(\frac{1}{2})}\frac{m_\tau}{n}\log\frac{n}{m_\tau}$. At last, use $h(\frac{1}{2}) \geq \frac{1}{5}$.

## REFERENCES

[1] Miller GA. Wordnet: a lexical database for english. *Communications of the ACM* **38** (1995) 39–41.

[2] McCray AT. An upper-level ontology for the biomedical domain. *Comparative and Functional Genomics* **4** (2003) 80–84.

[3] Gabrilovich E, Markovitch S. Wikipedia-based semantic interpretation for natural language processing. *Journal of Artificial Intelligence Research* **34** (2009) 443–498.

[4] Scott S, Matwin S. Feature engineering for text classification. *ICML* (1999), vol. 99, 379–388.

[5] Ferrucci D, Brown E, Chu-Carroll J, Fan J, Gondek D, Kalyanpur AA, et al. Building watson: An overview of the deepqa project. *AI magazine* **31** (2010) 59–79.

[6] Hoff PD, Raftery AE, Handcock MS. Latent space approaches to social network analysis. *Journal of the American Statistical Association* **97** (2002) 1090–1098.

[7] Bordes A, Usunier N, Garcia-Duran A, Weston J, Yakhnenko O. Translating embeddings for modeling multi-relational data. Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, editors, *Advances in Neural Information Processing Systems 26* (Curran Associates, Inc.) (2013), 2787–2795.

[8] Wang Z, Zhang J, Feng J, Chen Z. Knowledge graph embedding by translating on hyperplanes. *AAAI* (2014), 1112–1119.

[9] Yang B, Yih SWt, He X, Gao J, Deng L. Embedding entities and relations for learning and inference in knowledge bases. *Proceedings of the International Conference on Learning Representations* (2015).

[10] Lin Y, Liu Z, Sun M, Liu Y, Zhu X. Learning entity and relation embeddings for knowledge graph completion. *AAAI* (2015), 2181–2187.

[11] Garcia-Duran A, Bordes A, Usunier N, Grandvalet Y. Combining two and three-way embedding models for link prediction in knowledge bases. *Journal of Artificial Intelligence Research* **55** (2016) 715–742.

[12] Trouillon T, Welbl J, Riedel S, Gaussier É, Bouchard G. Complex embeddings for simple link prediction. *International Conference on Machine Learning* (2016), 2071–2080.

[13] Nickel M, Rosasco L, Poggio TA, et al. Holographic embeddings of knowledge graphs. *AAAI* (2016), 1955–1961.

[14] Liu H, Wu Y, Yang Y. Analogical inference for multi-relational embeddings. *Proceedings of the 34th International Conference on Machine Learning* (2017), vol. 70, 2168–2178.

[15] Socher R, Chen D, Manning CD, Ng A. Reasoning with neural tensor networks for knowledge base completion. *Advances in neural information processing systems* (2013), 926–934.

[16] Kotnis B, Nastase V. Analysis of the impact of negative sampling on link prediction in knowledge graphs. *arXiv preprint arXiv:1708.06816* (2017).

[17] Kanojia V, Maeda H, Togashi R, Fujita S. Enhancing knowledge graph embedding with probabilistic negative sampling. *Proceedings of the 26th International Conference on World Wide Web Companion* (2017), 801–802.

[18] Min B, Grishman R, Wan L, Wang C, Gondek D. Distant supervision for relation extraction with an incomplete knowledge base. *HLT-NAACL* (2013), 777–782.

[19] Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B* **67** (2005) 301–320.

[20] Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research* **12** (2011) 2121–2159.

[21] Robbins H, Monro S. A stochastic approximation method. *The Annals of Mathematical Statistics* **22** (1951) 400–407.

[22] Hitchcock FL. The expression of a tensor or a polyadic as a sum of products. *Journal of Mathematics and Physics* **6** (1927) 164–189.

[23] Nickel M, Tresp V, Kriegel HP. A three-way model for collective learning on multi-relational data. *Proceedings of the 28th International Conference on International Conference on Machine Learning* (Madison, WI, USA: Omnipress) (2011), ICML'11, 809–816.

[24] Tran N, Abramenko O, Jung A. On the sample complexity of graphical model selection from non-stationary samples. *IEEE Transactions on Signal Processing* **68** (2020) 17–32.

[25] Jung A, Hero, III AO, Mara AC, Jahromi S, Heimowitz A, Eldar YC. Semi-supervised learning in network-structured data via total variation minimization. *IEEE Transactions on Signal Processing* **67** (2019) 6256–6269.

[26] Massart P. *Concentration inequalities and model selection*, *Lecture notes in Mathematics*, vol. 1896 (Springer) (2007), 105–106 .

[27] Cover TM, Thomas JA. *Elements of Information Theory, Second Edition.* (Wiley) (2006), 37–41 .