Quantum Machine Learning on Knowledge Graphs

Yunpu Ma
LMU, Siemens

Volker Tresp
LMU, Siemens

Abstract

Knowledge Graphs (KGs) are large-scale triple-oriented relational databases for knowledge representation and reasoning. Implicit knowledge can be inferred by modeling and reconstructing the KGs. However, modeling becomes more and more computational resource intensive with the growing size of KGs. In this work we present the first quantum machine learning algorithm for knowledge graphs. This sampling-based quantum algorithm exhibits exponential acceleration w.r.t. the size of KGs during the inference task.

1 Introduction

KGs are graph-structured relational database consisting of semantic triples *(subject, predicate, object)*, where subject and object are nodes in the graph and predicate indicates the labeled arrow from the subject to the object. On the other hand, a knowledge graph can be seen as a tensor with three dimensions: one stands for subjects, one for predicates, and one for objects. Conventionally, we let $\chi \in \{0,1\}^{d_1 \times d_2 \times d_3}$ denote the KG semantic tensor, where d_1, d_2, and d_3 represents the number of subjects, predicates, and objects, respectively. An entry x_{spo} in χ takes value 1 if the semantic triple (s,p,o) is true, while it takes value 0 if the triple is simply wrong or missing. Values for the missing entries can be partly restored by modeling the observed entires. KGs can be modeled using tensor models, e.g., Tucker [20], PARAFAC [9], RESCAL [17], or compositional models, e.g., DistMult [22], HolE [16], HolNN [13].

In practice one might notice that inference tasks demand huge computing resources. This is because, given an incomplete semantic triple, say $(\mathrm{s}, \mathrm{p}, ?)$, the running time for inferring the correct objects to the query scales as $\mathcal{O}(d_3)$. The same algorithm has to be repeated at least d_3 times in order to determine possible answers leading to huge waste of computing power; especially, when nowadays the sizes of knowledge graphs are consistently growing. Thus the goal of this paper is to find quantum algorithms with potential acceleration to the inference tasks.

Quantum machine learning [3] is becoming an active research area by attracting researchers from different communities. It exhibits great potentials in speeding up classical algorithms, e.g., solving linear systems of equations [8], supervised and unsupervised learning [21], reinforcement learning [6], recommendation systems [10], etc. In this work, we present a quantum algorithm for modeling the knowledge graphs which shows exponential acceleration w.r.t. the size of knowledge graph. In particular, the knowledge graph is modeled by quantum singular value decomposition and projection, and the inference is achieved by sampling quantum states.

2 Tensor SVD

Firs, we recap singular value decomposition (SVD) of matrices. Then we introduce tensor SVD, and show that a given tensor can be reconstructed with small error from the low-rank tensor SVD of the subsampled tensor.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

SVD Let $A \in \mathbb{R}^{m\times n}$, the SVD is a factorization of A is the form $A = U\Sigma V^\intercal$, where Σ is a rectangle diagonal matrix singular values on the diagonal, $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ are orthogonal matrices with $U^\intercal U = UU^\intercal = I_m$ and $V^\intercal V = VV^\intercal = I_n$.

Notations for Tensors We adopt the notations from [4]. A N-way tensor is defined as $\mathcal{A} = (a_{i_1 i_2 \cdots i_N}) \in \mathbb{R}^{d_1\times d_2\times\cdots\times d_N}$, where d_n is the n-th dimension. Given two tensors $\mathcal{A}$ and $\mathcal{B}$ with the same dimensions, the inner product is defined as $\langle \mathcal{A}, \mathcal{B}\rangle_F := \sum_{i_N=1}^{d_N} \cdots \sum_{i_1=1}^{d_1} a_{i_1 i_2\cdots i_N} b_{i_1 i_2\cdots i_N}$. The Frobenius norm is defined as $||\mathcal{A}||_F := \sqrt{\langle \mathcal{A}, \mathcal{A}\rangle_F}$. The spectral norm of a tensor is defined as

$$||\mathcal{A}||_\sigma := \max\{\mathcal{A} \otimes_1 \mathbf{x}_1 \otimes \cdots \otimes_N \mathbf{x}_N | \mathbf{x}_k \in S^{d_k-1}, k = 1, \cdots, N\}, \tag{1}$$

where the tensor-vector product is defined as $\mathcal{A} \otimes_1 \mathbf{x}_1 \otimes \cdots \otimes_N \mathbf{x}_N := \sum_{i_N=1}^{d_N} \cdots \sum_{i_1=1}^{d_1} \mathcal{A}_{i_1 i_2\cdots i_N} x_{1 i_1} x_{2 i_2} \cdots x_{N i_N}$ and S^{d_k-1} is the unit sphere in $\mathbb{R}^{n_k}$.

Tensor SVD Parallel to the matrix singular value decomposition, *tensor singular value decomposition* was first studied in [4].

Definition 1. *[4] If a tensor $\mathcal{A} \in \mathbb{R}^{d_1\times d_2\times\cdots\times d_N}$ can be written as sum of rank-1 outer product tensors $\mathcal{A} = \sum_{i=1}^R \sigma_i u_1^{(i)} \otimes u_2^{(i)} \otimes \cdots \otimes u_N^{(i)}$, with singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_R > 0$ and $\langle u_k^{(i)}, u_k^{(j)}\rangle = \delta_{ij}$ for $k = 1, \cdots, N$. Then $\mathcal{A}$ has a* tensor singular value decomposition *with rank R.*

Define the orthogonal matrices $U_k = [u_k^{(1)}, u_k^{(2)}, \cdots, u_k^{(R)}] \in \mathbb{R}^{d_k\times R}$ with $U_k^T U_k = \mathbb{I}_R$ for $k = 1, \cdots, N$, and the diagonal tensor $\mathcal{D} \in \mathbb{R}^{R\times R\times\cdots\times R}$ with $\mathcal{D}_{ii\cdots i} = \sigma_i$, then the tensor SVD for $\mathcal{A}$ can be also written as $\mathcal{A} = \mathcal{D} \otimes_1 U_1 \otimes_2 U_2 \otimes \cdots \otimes_N U_N$.

Consider a given tensor $\mathcal{A}$, an interesting question is to find the low-rank tensor SVD of it.[4] proves the existence of the global optimum of the following optimization problem

$$\min ||\mathcal{A} - \sum_{i=1}^R \sigma_i u_1^{(i)} \otimes u_2^{(i)} \otimes \cdots \otimes u_N^{(i)}||_F \quad ; \quad \text{s.t.} \quad \langle u_k^{(i)}, u_k^{(j)}\rangle = \delta_{ij}, \text{ for } k = 1, \cdots, N$$

for an arbitrary $\mathcal{A} \in \mathbb{R}^{d_1\times d_2\times\cdots\times d_N}$ and $R \leq \min\{d_1, d_2, \cdots, d_N\}$ being the rank of tensor SVD.

Our quantum algorithm relies on the assumption that the semantic tensor χ can be approximated by a low rank tensor $\hat{\chi}$ with $||\chi - \hat{\chi}||_F^2 \leq \epsilon||\chi||_F^2$ for small $\epsilon > 0$. Previous work on recommendation systems [5] shows that the quality of recommendations for users depends on the reconstruction error. Similarly, in the case of relational learning, having a bounded tensor approximation error, it is possible to estimate the probability for a *bad* information retrieval. Consider a query $(\mathrm{s}, \mathrm{p}, ?)$. We normally only read top-n returns from the reconstructed tensor $\hat{\chi}$, written as $\hat{x}_{\mathrm{sp}1}, \cdots, \hat{x}_{\mathrm{sp}n}$, where n is a small integer and related to the commonly used Hits@n metric. The information retrieval is called *successful* if the correct answer to the query is in the list $\hat{x}_{\mathrm{sp}1}, \cdots, \hat{x}_{\mathrm{sp}n}$ which are assumed to be lager than a threshold δ. We have the following estimation.

Lemma 1. *If an algorithm returns an approximation of the semantic tensor χ, denoted $\hat{\chi}$, with $||\chi - \hat{\chi}||_F^2 \leq \epsilon||\chi||_F^2$, then the probability of an unsuccessful information retrieval from the top-n returns of $\hat{\chi}$ is bounded by $\frac{\epsilon}{n\delta^2}$. (Proof in A.1)*

In real-world applications we can only observe part of the non-zero entries in a given tensor $\mathcal{A}$, and the task is to infer unobserved non-zero entries with high probability. This corresponds to items recommendation for users given an observed preference matrix, or implicit knowledge inference given partially observed relational data. The partially observed tensor is called as subsampled or sparsified, denoted $\hat{\mathcal{A}}$. Matrix sparsification was first studied in [2], and tensor sparsification in [14]. Without further specifying the dimensionality of the tensor, we consider the following subsampling and rescaling scheme proposed in [2]:

$$\hat{\mathcal{A}}_{i_1 i_2\cdots i_N} = \begin{cases} \frac{\mathcal{A}_{i_1 i_2\cdots i_N}}{p} & \text{with probability } p \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

It means that the non-zero elements of a tensor are independently and identically sampled with the probability p and rescaled afterwards.

Now, the task is converted to reconstruct the original tensor $\mathcal{A}$ by modeling $\hat{\mathcal{A}}$. We will use tenor SVD to model the observed tensor $\hat{\mathcal{A}}$. The reconstruction error can be bounded either using the truncated r-rank tenor SVD, denoted $\hat{\mathcal{A}}_r$, or the projected tensor SVD with absolute singular value threshold τ, denoted $\hat{\mathcal{A}}_{|\cdot|\geq\tau}$. Theorem 1 and 2 give the bounds and the corresponding conditions on the sample probability. Some experimental results of classical tensor SVD are relegated to A.7.

Theorem 1. *Let $\mathcal{A} \in \{0,1\}^{d_1\times d_2\times\cdots\times d_N}$, which can be sufficiently approximated by its tensor SVD. Using the subsampling scheme defined in Eq. 2 with the sample probability $p = 8r\left(\log(\frac{2N}{N_0})\sum_{k=1}^{N} d_k + \log\frac{2}{\delta}\right)/(\tilde{\epsilon}||\mathcal{A}||_F)^2$, then the original tensor $\mathcal{A}$ can be reconstructed with bounded error $||\mathcal{A} - \hat{\mathcal{A}}_r||_F \leq \epsilon||\mathcal{A}||_F$ with probability at least $1-\delta$, where ϵ is a function of $\tilde{\epsilon}$. (Proof in A.2)*

Theorem 2. *Let $\mathcal{A} \in \{0,1\}^{d_1\times d_2\times\cdots\times d_N}$, which can be sufficiently approximated by its tensor SVD. Suppose $\hat{\mathcal{A}}$ is the sparsified tensor using the subsampling scheme defined in Eq. 2, then $\mathcal{A}$ can be reconstructed from the projected tensor SVD of $\hat{\mathcal{A}}$ with bounded error $||\mathcal{A} - \hat{\mathcal{A}}_{|\cdot|\geq\tau}||_F \leq \epsilon||\mathcal{A}||_F$ by carefully choosing the threshold τ and the sample probability p. (Proof in A.2)*

3 Quantum Machine Learning Algorithm for Knowledge Graphs

In this section we propose a quantum algorithm for inference on knowledge graphs using quantum singular value estimation. In the following we focus on the semantic tensor $\chi \in \{0,1\}^{d_1\times d_2\times d_3}$, and let $\hat{\chi}$ denote the partially observed part. Since knowledge graphs contain significant global relational patterns [15], χ can be reconstructed sufficiently by $\hat{\chi}$ according to Theorem 1.

Moreover, w.l.o.g., we consider querying on the correct objects given $(\mathrm{s}, \mathrm{p}, ?)$. Recall that in the case of recommendation system a preference matrix could have multiple nonzero entries in a given user-row, and recommendations are made according to nonzero entries by assuming that the user is 'typical' [5]. However, in a KG there might be only one nonzero entry in the row $(\mathrm{s}, \mathrm{p}, \cdot)$. Thus, for inference on the KG quantum algorithm needs to sample triples with the given subject s and post-select on the predicate p. This can be a valid step if the number of semantic triples having s as subject is $\mathcal{O}(1)$.

The most technical challenge in quantum machine learning is to load classical data into quantum registers or states, since reading or writing high-dimensional data might directly destroy the quantum acceleration gained with respect to the dimensionality of the data. Thus, a technique quantum Random Access Memory (qRAM) [7] was developed, which can map a classical data vector into its quantum state with exponential acceleration. A.3 shows that there exits a classical memory structure for implementing qRAM.

We briefly sketch the quantum algorithm. The basic idea is to project the observed data onto the space spanned by the eigenspaces of $\hat{\chi}$ whose corresponding singular values are larger than a threshold. Thus, we need to create an operator which can reveal the eigenspaces of $\hat{\chi}$, and a quantum algorithm for estimating the singular values.

The first step is to prepare the following quantum state from $\hat{\chi}$:

$$\rho_{\hat{\chi}^\dagger\hat{\chi}} := \sum_{i_2 i_3 i_2' i_3'} \mathcal{C}_{i_2 i_3 i_2' i_3'} |i_2 i_3\rangle \langle i_2' i_3'| = \sum_{i_2 i_3 i_2' i_3'} \sum_{i_1} \hat{\chi}^\dagger_{i_1,i_2 i_3}\hat{\chi}_{i_1,i_2' i_3'} |i_2 i_3\rangle \langle i_2' i_3'|,$$

where $\sum_{i_1} \hat{\chi}^\dagger_{i_1,i_2 i_3}\hat{\chi}_{i_1,i_2' i_3'}$ means tenor contraction along the first dimension.

Lemma 2. *$\rho_{\hat{\chi}^\dagger\hat{\chi}}$ can be prepared via qRAM in time $\mathcal{O}(\log(d_1 d_2 d_3))$ (Proof see A.4).*

Since we assume that the subsampled tensors $\hat{\chi}$ can be sufficiently approximated by its tensor SVD, namely $\hat{\chi} \approx \sum_{i=1}^{R} \sigma_i u_1^{(i)} \otimes u_2^{(i)} \otimes u_3^{(i)}$, the density $\rho_{\hat{\chi}^\dagger\hat{\chi}}$ will be decomposed as $\rho_{\hat{\chi}^\dagger\hat{\chi}} = \frac{1}{\sum_{i=1}^{R}\sigma_i^2}\sum_{i=1}^{R}\sigma_i^2 |u_i^{(2)}\rangle \otimes |u_i^{(3)}\rangle \langle u_i^{(2)}| \otimes \langle u_i^{(3)}|$. In [12] this is called quantum state self-tomography.

The next step is to estimate singular values of $\rho_{\hat{\chi}^\dagger\hat{\chi}}$ via the method proposed in [12] which is referred to as quantum principal component analysis (qPCA). The key is to prepare the unitary operator

$U = \sum_{k=0}^{K-1} |k\ \Delta t\rangle \langle k\ \Delta t|_C \otimes \exp(-ik\Delta t \tilde{\rho}_{\hat{\chi}^\dagger \hat{\chi}})$, where the clock register C is needed for the phase estimation, $K\Delta t$ determines the precision of estimated singular values, and $\tilde{\rho}_{\hat{\chi}^\dagger \hat{\chi}}$ is the rescaled density matrix. The following Lemma shows that the unitary operator $\mathrm{e}^{-it\tilde{\rho}_{\hat{\chi}^\dagger \hat{\chi}}}$ can be applied on an arbitrary quantum state for any given t.

Lemma 3. *[19] Unitary operator $\mathrm{e}^{-it\tilde{\rho}_{\hat{\chi}^\dagger \hat{\chi}}}$ can be applied to any quantum state, where $\tilde{\rho}_{\hat{\chi}^\dagger \hat{\chi}} := \frac{\rho_{\hat{\chi}^\dagger \hat{\chi}}}{d_2 d_3}$. The total run time of simulation is $\mathcal{O}(\frac{t^2}{\epsilon} T_\rho)$, where ϵ is the accuracy, and T_ρ is the time for accessing the density matrix and simulating the unitary operator on quantum state. (see A.5)*

Since we sample triples given the subject s, a quantum state $|\hat{\chi}^{(1)}_{\mathrm{s}}\rangle_I$ needs to be created first, where $\hat{\chi}^{(1)}_{\mathrm{s}}$ denotes the s-row of the flattened tensor $\hat{\chi}$ along the first dimension, and I indicates the input register. Afterwards the operator U is applied to the quantum state $\sum_{k=0}^{K-1} |k\Delta t\rangle_C \otimes |\hat{\chi}^{(1)}_{\mathrm{s}}\rangle_I$. After this stage of computation, we obtain

$$\sum_{i=1}^{R} \beta_i \left(\sum_{k=0}^{K-1} \mathrm{e}^{-ik\ \Delta t\ \tilde{\sigma}_i^2}\ |k\ \Delta t\rangle_C \right) |u_i^{(2)}\rangle_I \otimes |u_i^{(3)}\rangle_I \,, \tag{3}$$

where $\tilde{\sigma}_i := \frac{\sigma_i}{\sqrt{d_2 d_3}}$ are the rescaled singular values of $\tilde{\rho}_{\hat{\chi}^\dagger \hat{\chi}}$. Moreover, β_i are the coefficients of $|\hat{\chi}^{(1)}_{\mathrm{s}}\rangle_I$ written in the basis $|u_2^{(i)}\rangle_I \otimes |u_3^{(i)}\rangle_I$, namely $|\hat{\chi}^{(1)}_{\mathrm{s}}\rangle_I = \sum_{i=1}^{R} \beta_i |u_2^{(i)}\rangle_I \otimes |u_3^{(i)}\rangle_I$.

The third step is to perform the quantum phase estimation algorithm on the clock register C. The quantum phase estimation was first propose in [11] and given in the A.6. The resulting state reads $\sum_{i=1}^{R} \beta_i |\lambda_i\rangle_C \otimes |u_2^{(i)}\rangle_I \otimes |u_3^{(i)}\rangle_I$ where $\lambda_i := \frac{2\pi}{\tilde{\sigma}_i^2}$. This step can be understood as follows: The probability amplitude of measuring the register C is maximized when $k\ \Delta t = \frac{2\pi}{\tilde{\sigma}_i^2}$ (see the quantum state in Eq. 3). Thus, the time step Δt determines the accuracy of quantum phase estimation. We chose $\Delta t = \mathcal{O}(\frac{1}{\epsilon})$, and according to Lemma 3 the total run time is $\mathcal{O}(\frac{1}{\epsilon^3} T_\rho) = \mathcal{O}(\frac{1}{\epsilon^3} \mathrm{polylog}(d_1 d_2 d_3))$. We can also perform controlled computation on the register to recover the original singular values, and obtain $\sum_{i=1}^{R} \beta_i |\sigma_i^2\rangle_C \otimes |u_2^{(i)}\rangle_I \otimes |u_3^{(i)}\rangle_I$.

The next step is to perform quantum singular values projection on the quantum state from the last step. Classically, this step corresponds to projecting $\hat{\chi}$ onto $\hat{\chi}_{|\cdot|\geq\tau}$. In this way, observed signals will be smoothed and unobserved signals can be boosted from which we can infer unseen triples $(\mathrm{s}, \mathrm{p}, ?)$ in the test dataset (see Theorem 2). The quantum projection given the threshold $\tau > 0$ can be performed in the following way: Create a new register R and a unitary operator that maps $|\sigma_i^2\rangle_C \otimes |0\rangle_R$ to $|\sigma_i^2\rangle_C \otimes |1\rangle_R$ if $\sigma_i^2 < \tau^2$. This step of projection gives

$$\sum_{i:\sigma_i^2 \geq \tau^2} \beta_i |\sigma_i^2\rangle_C \otimes |u_2^{(i)}\rangle_I \otimes |u_3^{(i)}\rangle_I \otimes |0\rangle_R + \sum_{i:\sigma_i^2 < \tau^2} \beta_i |\sigma_i^2\rangle_C \otimes |u_2^{(i)}\rangle_I \otimes |u_3^{(i)}\rangle_I \otimes |1\rangle_R \,. \tag{4}$$

The last step is to erase the clock register, and measure be new register R and post-select the state $|0\rangle_R$. This gives the state $\sum_{i:\sigma_i^2 \geq \tau^2} \beta_i |u_2^{(i)}\rangle_I \otimes |u_3^{(i)}\rangle_I = |\hat{\chi}^+_{|\cdot|\geq\tau} \hat{\chi}_{|\cdot|\geq\tau} \hat{\chi}^{(1)}_{\mathrm{s}}\rangle$, where $\hat{\chi}^+_{|\cdot|\geq\tau} \hat{\chi}_{|\cdot|\geq\tau} := \sum_{i:\sigma_i^2 \geq \tau^2} (u_2^{(i)} \otimes u_3^{(i)}) \otimes (u_2^{(i)} \otimes u_3^{(i)})$ which is similar to the pseudoinverse in the case of matrices. Finally, we can measure this state in the standard the basis to get the triples with subject s, and post-select on the predicate p. This will return objects to the inference $(\mathrm{s}, \mathrm{p}, ?)$ after $\mathcal{O}(\frac{1}{\epsilon^3} \mathrm{polylog}(d_1 d_2 d_3))$ times of repetitions. The quantum algorithm is summarized in A.8.

4. Conclusion In this work we present a quantum machine learning algorithm showing exponentially accelerated inference on knowledge graphs. We first prove that the semantic tensor can be reconstructed from the projected tensor SVD of the subsampled tensor with small error. Afterwards, we construct the quantum algorithm using quantum singular value estimation and projection. The resulting sample-based quantum machine learning algorithm shows an exponential acceleration w.r.t. the dimensions of the semantic tensor. However, the proposed quantum algorithm cannot be fully implemented on the state-of-the-art quantum computers due to the limited number of fully entangled qubits [18] and the technical difficulties of implementing qRAM [1].

References

[1] Scott Aaronson. Read the fine print. *Nature Physics*, 11(4):291, 2015.

[2] Dimitris Achlioptas and Frank McSherry. Fast computation of low-rank matrix approximations. *Journal of the ACM (JACM)*, 54(2):9, 2007.

[3] Jacob Biamonte, Peter Wittek, Nicola Pancotti, Patrick Rebentrost, Nathan Wiebe, and Seth Lloyd. Quantum machine learning. *Nature*, 549(7671):195, 2017.

[4] Jie Chen and Yousef Saad. On the tensor svd and the optimal low rank orthogonal approximation of tensors. *SIAM Journal on Matrix Analysis and Applications*, 30(4):1709–1734, 2009.

[5] Petros Drineas, Iordanis Kerenidis, and Prabhakar Raghavan. Competitive recommendation systems. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 82–90. ACM, 2002.

[6] Vedran Dunjko, Jacob M Taylor, and Hans J Briegel. Quantum-enhanced machine learning. *Physical review letters*, 117(13):130501, 2016.

[7] Vittorio Giovannetti, Seth Lloyd, and Lorenzo Maccone. Quantum random access memory. *Physical review letters*, 100(16):160501, 2008.

[8] Aram W Harrow, Avinatan Hassidim, and Seth Lloyd. Quantum algorithm for linear systems of equations. *Physical review letters*, 103(15):150502, 2009.

[9] Richard A Harshman. Foundations of the parafac procedure: Models and conditions for an" explanatory" multimodal factor analysis. 1970.

[10] Iordanis Kerenidis and Anupam Prakash. Quantum recommendation systems. *arXiv preprint arXiv:1603.08675*, 2016.

[11] A Yu Kitaev. Quantum measurements and the abelian stabilizer problem. *arXiv preprint quant-ph/9511026*, 1995.

[12] Seth Lloyd, Masoud Mohseni, and Patrick Rebentrost. Quantum principal component analysis. *Nature Physics*, 10(9):631, 2014.

[13] Yunpu Ma, Marcel Hildebrandt, Stephan Baier, and Volker Tresp. Holistic representations for memorization and inference.

[14] Nam H Nguyen, Petros Drineas, and Trac D Tran. Tensor sparsification via a bound on the spectral norm of random tensors. *arXiv preprint arXiv:1005.4732*, 2010.

[15] Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33, 2016.

[16] Maximilian Nickel, Lorenzo Rosasco, Tomaso A Poggio, et al. Holographic embeddings of knowledge graphs. In *AAAI*, volume 2, pages 3–2, 2016.

[17] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. A three-way model for collective learning on multi-relational data. In *ICML*, volume 11, pages 809–816, 2011.

[18] John Preskill. Quantum computing in the nisq era and beyond. *arXiv preprint arXiv:1801.00862*, 2018.

[19] Patrick Rebentrost, Adrian Steffens, Iman Marvian, and Seth Lloyd. Quantum singular-value decomposition of nonsparse low-rank matrices. *Physical review A*, 97(1):012327, 2018.

[20] Ledyard R Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966.

[21] Nathan Wiebe, Ashish Kapoor, and Krysta Svore. Quantum algorithms for nearest-neighbor methods for supervised and unsupervised learning. *arXiv preprint arXiv:1401.2142*, 2014.

[22] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*, 2014.

A Appendix

A.1 Proof of Lemma 1

Proof. Without loss of generality, consider the retrieval of objects given the inference task $(\mathrm{s}, \mathrm{p}, ?)$. The retrieval becomes unsuccessful if the correct objects regarding to the query cannot be found in the top-n returns from $\hat{\chi}$. Suppose the entries of χ are 0 and 1, and $\hat{x}_{\mathrm{sp1}} \geq \cdots \geq \hat{x}_{\mathrm{sp}n} \geq \delta$. Then the minimum contribution of an unsuccessful retrieval to the reconstruction error $||\chi - \hat{\chi}||_F^2$ is $n\delta^2 + 1 \approx n\delta^2$. Ignore the multiplicity of the correct answers to a single query, and since the reconstruction error is bounded by $\epsilon||\chi||_F^2$, the probability of an unsuccessful retrieval is bounded by $\frac{\epsilon}{n\delta^2}$. ∎

A.2 Proof of Theorem 1 and Theorem 2

We introduce the following notations to bound the reconstruction error from the subsampled tensor. Consider a N-way tensor $\mathcal{A} \in \mathbb{R}^{d_1 \times d_2 \times \cdots \times d_N}$, which has a tensor SVD with full rank R. Let $\mathcal{A}_r = \mathcal{D} \otimes_1 U_1 \otimes_2 U_2 \otimes \cdots \otimes_N U_N$ denote the truncated r-rank tenor SVD of $\mathcal{A}$ with $U_i \in \mathbb{R}^{d_i \times r}$ for $i = 1, \cdots, N$. Define the projection operators $\mathcal{P}_i^{\mathcal{A},r} := \mathbb{I} \otimes \otimes \cdots \otimes U_i U_i^T \otimes \cdots \otimes \mathbb{I}$ with $i = 1, \cdots, N$, and the product projections $\mathcal{P}^{\mathcal{A},r} := \prod_{i=1}^N \mathcal{P}_i^{\mathcal{A},r}$. We have:

Lemma A 1. $\mathcal{P}^{\mathcal{A},r}\mathcal{A} = \mathcal{A}_r$.

Proof. Let $\mathcal{A}_R = \tilde{\mathcal{D}} \otimes_1 \tilde{U}_1 \otimes \cdots \otimes \tilde{U}_N$ denotes the full rank tensor SVD of $\mathcal{A}$, where $\tilde{U}_i = [u_i^{(1)}, u_i^{(2)}, \cdots, u_i^{(R)}]$ for $i = 1, \cdots, N$. Define $\mathcal{A}_R^{\perp} := \mathcal{A} - \mathcal{A}_R$, then we have $\langle \mathcal{A}_R^{\perp}, \mathcal{T}_i \rangle = 0$ with $\mathcal{T}_i := u_1^{(i)} \otimes u_2^{(i)} \otimes \cdots \otimes u_N^{(i)}$ for $i = 1, \cdots, R$. To see this, suppose $\exists j$, such that $\langle \mathcal{A}_R^{\perp}, \mathcal{T}_j \rangle = \epsilon \neq 0$. Then,

$$||\mathcal{A} - \sum_{i=1}^{R} \sigma_i \mathcal{T}_i - \epsilon \mathcal{T}_j||_F^2 = ||\mathcal{A} - \sum_{i=1}^{R} \sigma_i \mathcal{T}_i||_F^2 - \epsilon^2 < ||\mathcal{A} - \sum_{i=1}^{R} \sigma_i \mathcal{T}_i||_F^2,$$

which contradicts the fact that $\mathcal{A}_R$ is the global minimum of the optimization. Thus, $\mathcal{P}^{\mathcal{A},r}\mathcal{A} = \mathcal{P}^{\mathcal{A},r}(\mathcal{A}_R + \mathcal{A}_R^{\perp}) = \prod_{i=1}^N \mathcal{P}_i^{\mathcal{A},r}\mathcal{A}_R = \mathcal{A}_r$. ∎

Consider two tensors $\mathcal{A}$ and $\mathcal{B}$ which can be sufficiently approximated by their tensor SVDs, and their r-rank projection operators $\mathcal{P}^{\mathcal{A},r}$ and $\mathcal{P}^{\mathcal{B},r}$. We can derive the inequality $||\mathcal{P}^{\mathcal{A},r}\mathcal{A}||_F \geq ||\mathcal{P}^{\mathcal{B},r}\mathcal{A}||_F$. Parallel to Lemma 4 in [1], we can have the following bound.

Lemma A 2. *[1] Given tensors $\mathcal{A}$, $\mathcal{B}$ and their r-rank tensor SVD approximations $\mathcal{A}_r = \mathcal{P}^{\mathcal{A},r}\mathcal{A}$, $\mathcal{B}_r = \mathcal{P}^{\mathcal{B},r}\mathcal{B}$, we have*

$$||\mathcal{A} - \mathcal{B}_r||_F \leq ||\mathcal{A} - \mathcal{A}_r||_F + 2\sqrt{||(\mathcal{A} - \mathcal{B})_r||_F ||\mathcal{A}_r||_F} + ||(\mathcal{A} - \mathcal{B})_r||_F. \tag{1}$$

Consider a tensor $\mathcal{A}$ that is subsampled and rescaled. This perturbed tensor can be written as $\hat{\mathcal{A}} = \mathcal{A} + \mathcal{N}$, where $\mathcal{N}$ is a random tensor. In the following, we use $\hat{\mathcal{A}}$ to represent subsampled (sparsified) tensor, and $\hat{\mathcal{A}}_r$ the truncated r-rank tensor SVD of $\hat{\mathcal{A}}$. Thus, according to Lemma 2, the reconstruction error from the truncated $\hat{\mathcal{A}}$ is bounded by

$$||\mathcal{A} - \hat{\mathcal{A}}_r||_F \leq ||\mathcal{A} - \mathcal{A}_r||_F + 2\sqrt{||\mathcal{N}_r||_F ||\mathcal{A}_r||_F} + ||\mathcal{N}_r||_F. \tag{2}$$

In order to further estimate the bound of the error, we briefly recap the tensor subsampling and sparsification techniques. The basic idea behind matrix/tensor sparsification algorithms is to neglect all small entries, and keep or amplify sufficiently large entries, such that the original matrix/tensor can be reconstructed element-wise with bounded error. Matrix sparsification was first studied in [1], and tensor sparsification in [5]. Recall that the semantic tensor contains entries 0 and 1. Thus, sparsification of the semantic tensor is equivalent to the separation of training and test datasets, and the reconstruction is equivalent to the inference on the test dataset.

Without further specification, we consider the following general sparsification and rescaling method used in [1]:

$$\hat{\mathcal{A}}_{i_1 i_2 \cdots i_N} = \begin{cases} \frac{\mathcal{A}_{i_1 i_2 \cdots i_N}}{p} & \text{with probability } p \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where the choose of the element-wise sample probability p will be discussed later. Note that the expectation values of the entries of the sparsified tensor read $\mathbb{E}[\hat{\mathcal{A}}_{i_1 i_2 \cdots i_N}] = \mathcal{A}_{i_1 i_2 \cdots i_N}$. Recall that the perturbation is defined as $\mathcal{N} = \hat{\mathcal{A}} - \mathcal{A}$. Thus, the entries of the noise tensor have zero mean $\mathbb{E}[\mathcal{N}_{i_1 i_2 \cdots i_N}] = 0$ and variance $\mathrm{Var}[\mathcal{N}_{i_1 i_2 \cdots i_N}] = \mathcal{A}_{i_1 i_2 \cdots i_N}(\frac{1}{p} - 1)$.

We give the bounds of norms of the noise tensor $\mathcal{N}$, the proof closely follows [10].

Lemma A 3. *Assume that the noise tensor $\mathcal{N}$ is generated by subsampling a binary tensor $\mathcal{A} \in \{0,1\}^{d_1 \times d_2 \times \cdots \times d_N}$ according to Eq. 3. The spectral norm of $\mathcal{N}$ is bounded by*

$$||\mathcal{N}||_\sigma \leq \sqrt{\frac{8}{p}\left(\log(\frac{2N}{N_0})\sum_{k=1}^{N} d_k + \log\frac{2}{\delta}\right)}, \tag{4}$$

with probability at least $1 - \delta$.

Proof. Note that the entries $\mathcal{N}_{i_1 i_2 \cdots i_N}$ are independent having zero mean and bounded variance $\mathrm{Var}[\mathcal{N}_{i_1 i_2 \cdots i_N}] \leq \frac{1}{p} - 1$. We first estimate the following quantity

$$\begin{aligned} \ln \mathbb{E}[\mathrm{e}^{\eta \mathcal{N}_{i_1 i_2 \cdots i_N}}] &= \ln \mathbb{E}[\mathrm{e}^{\eta(\hat{\mathcal{A}}_{i_1 i_2 \cdots i_N} - \mathcal{A}_{i_1 i_2 \cdots i_N})}] \\ &= \ln\left[\mathrm{e}^{-\mathcal{A}_{i_1 i_2 \cdots i_N}\eta}\left(p\mathrm{e}^{\eta \mathcal{A}_{i_1 i_2 \cdots i_N}/p} + (1-p)\right)\right]. \end{aligned}$$

A series expansion around $\eta \approx 0$ reveals that

$$\ln \mathbb{E}[\mathrm{e}^{\eta \mathcal{N}_{i_1 i_2 \cdots i_N}}] \approx \eta^2 \frac{\mathcal{A}_{i_1 i_2 \cdots i_N}(1-p)}{2p} \leq \eta^2 \frac{1}{2p} \quad \Rightarrow \quad \mathbb{E}[\mathrm{e}^{\eta \mathcal{N}_{i_1 i_2 \cdots i_N}}] \leq \mathrm{e}^{\eta^2/2p} \tag{5}$$

by using the fact $\mathcal{A}_{i_1 i_2 \cdots i_N} \in \{0,1\}$.

Afterwards, the tensor-vector product $\mathcal{N} \otimes_1 \mathbf{x}_1 \cdots \otimes_N \mathbf{x}_N$ with $\mathbf{x}_k \in S^{d_k - 1}$, $k = 1, \cdots, N$ can be estimated by bounding the probability $\Pr(|\mathcal{N} \otimes_1 \mathbf{x}_1 \cdots \otimes_N \mathbf{x}_N| \geq t)$ for non-negative t. Given Eq. 5 we have

$$\mathbb{E}[\mathrm{e}^{s\mathcal{N}_{i_1 i_2 \cdots i_N} x_{1 i_1} x_{2 i_2} \cdots x_{N i_N}}] \leq \mathrm{e}^{s^2 x_{1 i_1}^2 x_{2 i_2}^2 \cdots x_{N i_N}^2 / 2p},$$

where s is an auxiliary variable. This gives

$$\begin{aligned} \Pr(\mathcal{N} \otimes_1 \mathbf{x}_1 \cdots \otimes_N \mathbf{x}_N \geq t) &= \Pr(\mathrm{e}^{s\mathcal{N} \otimes_1 \mathbf{x}_1 \cdots \otimes_N \mathbf{x}_N} \geq \mathrm{e}^{st}) \\ &\leq \mathrm{e}^{-st} \mathbb{E}[\mathrm{e}^{s\mathcal{N} \otimes_1 \mathbf{x}_1 \cdots \otimes_N \mathbf{x}_N}] \\ &\leq \exp\{-st + \frac{s^2}{2p} \sum_{i_1=1}^{d_1} \cdots \sum_{i_N=1}^{d_N} x_{1 i_1}^2 \cdots x_{N i_N}^2\} \\ &= \mathrm{e}^{-st + \frac{s^2}{2p}}. \end{aligned}$$

The above equation takes the minimum when $s = tp$, with $\Pr(\mathcal{N} \otimes_1 \mathbf{x}_1 \cdots \otimes_N \mathbf{x}_N \geq t) \leq \mathrm{e}^{-\frac{t^2 p}{2}}$. Similarly we have the probability $\Pr(\mathcal{N} \otimes_1 \mathbf{x}_1 \cdots \otimes_N \mathbf{x}_N \leq -t) \leq \mathrm{e}^{-\frac{t^2 p}{2}}$. Thus, in summary $\Pr(|\mathcal{N} \otimes_1 \mathbf{x}_1 \cdots \otimes_N \mathbf{x}_N| \geq t) \leq 2\mathrm{e}^{-\frac{t^2 p}{2}}$.

Using the covering number on unit spheres and the compactness of the space $S^{d_1 - 1} \times S^{d_2 - 1} \times \cdots \times S^{d_N - 1}$, the spectral norm of the noise tensor $\mathcal{N}$ can be bounded as follows:

$$||\mathcal{N}||_\sigma \leq \sqrt{\frac{8}{p}\left(\log(\frac{2N}{N_0})\sum_{k=1}^{N} d_k + \log\frac{2}{\delta}\right)} \tag{6}$$

with probability at least $1 - \delta$, where the constant $N_0 := \log\frac{3}{2}$ (see the proof of Theorem 2 in [10]). ∎

Using the facts that $||\mathcal{N}_r||_\sigma = ||\mathcal{N}||_\sigma$, and $||\mathcal{N}_r||_F \leq \sqrt{r}||\mathcal{N}_r||_\sigma$. We can estimate the norms of the truncated tensor SVD of the noise tensor.

Lemma A 4.

$$||\mathcal{N}_r||_\sigma \leq \sqrt{\frac{8}{p}\left(\log(\frac{2N}{N_0})\sum_{k=1}^{N} d_k + \log\frac{2}{\delta}\right)}$$
$$||\mathcal{N}_r||_F \leq \sqrt{r\frac{8}{p}\left(\log(\frac{2N}{N_0})\sum_{k=1}^{N} d_k + \log\frac{2}{\delta}\right)}.$$

Now we are able to determine the sample probability, such that the error ratio $\frac{||\mathcal{A}-\hat{\mathcal{A}}_r||_F}{||\mathcal{A}||_F}$ is bounded.

Theorem A 1 (Theorem 1 in the main text)**.** *Let $\mathcal{A} \in \{0,1\}^{d_1\times d_2\times\cdots\times d_N}$, which can be sufficiently approximated by its tensor SVD. Using the subsampling scheme defined in Eq. 3 with the sample probability $p = 8r\left(\log(\frac{2N}{N_0})\sum_{k=1}^{N} d_k + \log\frac{2}{\delta}\right)/(\tilde{\epsilon}||\mathcal{A}||_F)^2$, then the original tensor $\mathcal{A}$ can be reconstructed with bounded error $||\mathcal{A} - \hat{\mathcal{A}}_r||_F \leq \epsilon||\mathcal{A}||_F$ with probability at least $1-\delta$, where ϵ is a function of $\tilde{\epsilon}$.*

Proof. Suppose tensor $\mathcal{A}$ can be sufficiently approximated by its r-rank tensor SVD, in a sense that $||\mathcal{A} - \mathcal{A}_r|| \leq \epsilon_0||\mathcal{A}||_F$ for some small $\epsilon_0 > 0$. Let the Frobenius norm of the low-rank noise tensor $\mathcal{N}_r$ be bounded by $\tilde{\epsilon}||\mathcal{A}||_F$ with $\tilde{\epsilon} > 0$. According to Lemma A 4 the sample probability should satisfy $p \geq \frac{8r\left(\log(\frac{2N}{N_0})\sum_{k=1}^{N} d_k+\log\frac{2}{\delta}\right)}{(\tilde{\epsilon}||\mathcal{A}||_F)^2}$. Using Eq. 2 we have

$$||\mathcal{A} - \hat{\mathcal{A}}_r||_F \leq \epsilon_0||\mathcal{A}||_F + 2\sqrt{\tilde{\epsilon}}||\mathcal{A}||_F + \tilde{\epsilon}||\mathcal{A}||_F = \epsilon||\mathcal{A}||_F,$$

where $\epsilon := \epsilon_0 + 2\sqrt{\tilde{\epsilon}} + \tilde{\epsilon}$. ∎

Note that in the case where $\mathcal{A}$ is a two-dimensional matrix, the sample probability derived in [1] reads $\mathcal{O}(\frac{d_1+d_2}{||\mathcal{A}||_F^2})$. This corresponds the high-dimensional tensor case.

For the later use in the quantum algorithm, instead of considering low-rank approximation of the subsampled tensor, we study the tensor SVD with projected singular values, denoted as $\hat{\mathcal{A}}_{|\cdot|\geq\tau}$. This notation denotes that subsampled tensor $\hat{\mathcal{A}}$ is projected onto the eigenspaces with absolute singular values larger than a threshold. Later, it will be also referred to as the projected tensor SVD of $\hat{\mathcal{A}}$ with threshold τ. The following theorem discusses the choice of sample probability and threshold τ, such that the error ratio $\frac{||\mathcal{A}-\hat{\mathcal{A}}_{|\cdot|\geq\tau}||_F}{||\mathcal{A}||_F}$ is bounded.

Theorem A 2 (Theorem 2 in the main text)**.** *Let $\mathcal{A} \in \{0,1\}^{d_1\times d_2\times\cdots\times d_N}$, which can be sufficiently approximated by its tensor SVD. Suppose $\hat{\mathcal{A}}$ is the sparsified tensor using the subsampling scheme defined in Eq. 3, then $\mathcal{A}$ can be reconstructed from the projected tensor SVD of $\hat{\mathcal{A}}$ with bounded error $||\mathcal{A} - \hat{\mathcal{A}}_{|\cdot|\geq\tau}||_F \leq \epsilon||\mathcal{A}||_F$ by carefully choosing the threshold τ and the sample probability p.*

Proof. Suppose tensor $\mathcal{A}$ can be sufficiently approximated by its r-rank tensor SVD, in a sense that $||\mathcal{A} - \mathcal{A}_r|| \leq \epsilon_0||\mathcal{A}||_F$ for some small $\epsilon_0 > 0$. Define the threshold as $\tau := \kappa||\hat{\mathcal{A}}||_F$ for some $\kappa > 0$.

Let l_1 denote the largest index of singular values with $\sigma_{l_1} \geq \kappa||\hat{\mathcal{A}}||_F$, and let l_2 denote the smallest index of singular values with $\sigma_{l_2} \leq -\kappa||\hat{\mathcal{A}}||_F$. If the threshold τ is large enough, we consider the case $l_1 \ll l_2$. In addition, we have the following constrain

$$l_1 \cdot \sigma_{l_1}^2 \leq ||\hat{\mathcal{A}}_{l_1}||_F^2 \leq ||\hat{\mathcal{A}}||_F^2 \quad \Rightarrow \quad l_1 \cdot \kappa^2 \leq 1. \tag{7}$$

Recall that the full rank tensor SVD of $\hat{\mathcal{A}}$ is written as $\hat{\mathcal{A}}_R$, where the full rank R can be much larger than r. We first bound $||\mathcal{A} - \hat{\mathcal{A}}_{|\cdot|\geq\tau}||_F$ as follows:

$$\begin{aligned}
||\mathcal{A} - \hat{\mathcal{A}}_{|\cdot|\geq\tau}||_F &= ||\mathcal{A} - \hat{\mathcal{A}}_{[0,l_1]\cup[l_2,R]}||_F = ||\mathcal{A} - (\hat{\mathcal{A}}_R - \hat{\mathcal{A}}_{l_2} + \hat{\mathcal{A}}_{l_1})||_F \\
&\leq ||\mathcal{A} - \hat{\mathcal{A}}_{l_1}||_F + ||\hat{\mathcal{A}}_{l_2} - \hat{\mathcal{A}}_R||_F = ||\mathcal{A} - \hat{\mathcal{A}}_{l_1}||_F + ||\mathcal{A} - \mathcal{A} + \hat{\mathcal{A}}_{l_2} - \hat{\mathcal{A}}_R||_F \\
&\leq ||\mathcal{A} - \hat{\mathcal{A}}_{l_1}||_F + ||\mathcal{A} - \hat{\mathcal{A}}_R||_F + ||\mathcal{A} - \hat{\mathcal{A}}_{l_2}||_F \\
&\leq 3||\mathcal{A} - \hat{\mathcal{A}}_{l_1}||_F.
\end{aligned}$$

Assume $l_1 \ll l_2$, we only distinguish two cases: $l_2 \gg l_1 \geq r$ and $l_1 < r \ll l_2$.

If $l_1 \geq r$, we have

$$\begin{aligned}
||\mathcal{A} - \hat{\mathcal{A}}_{|\cdot|\geq\tau}||_F \leq 3||\mathcal{A} - \hat{\mathcal{A}}_{l_1}||_F &\leq 3(||\mathcal{A} - \mathcal{A}_{l_1}||_F + 2\sqrt{||\mathcal{N}_{l_1}||_F||\mathcal{A}||_F} + ||\mathcal{N}_{l_1}||_F) \\
&\leq 3(||\mathcal{A} - \mathcal{A}_r||_F + 2\sqrt{||\mathcal{N}_{l_1}||_F||\mathcal{A}||_F} + ||\mathcal{N}_{l_1}||_F).
\end{aligned}$$

Let $||\mathcal{N}_{l_1}||_F \leq \tilde{\epsilon}||\mathcal{A}||_F$ for some small $\tilde{\epsilon} > 0$. According to the Lemma 4 the sample probability should satisfy $p \geq \frac{l_1\, C_0}{(\tilde{\epsilon}||\mathcal{A}||_F)^2} := p_1$ where the constant is defined as $C_0 := 8\left(\log(\frac{2N}{N_0})\sum_{k=1}^{N} d_k + \log\frac{2}{\delta}\right)$. In this case $||\mathcal{A} - \hat{\mathcal{A}}_{|\cdot|\geq\tau}||_F \leq 3(\epsilon_0 + 2\sqrt{\tilde{\epsilon}} + \tilde{\epsilon})||\mathcal{A}||_F$ for $l_1 \geq r$.

On the other hand if $l_1 < r \ll l_2$, we first fix the sample probability $p = p_1$ and use the fact that $||\hat{\mathcal{A}}||_F \leq \sqrt{\frac{2}{p}}||\mathcal{A}||_F$ is satisfied with high probability (see the proof of Theorem 4.2 in [3]). It gives

$$\begin{aligned}
||\mathcal{A} - \hat{\mathcal{A}}_{|\cdot|\geq\sigma}||_F &\leq 3||\mathcal{A} - \hat{\mathcal{A}}_{l_1}||_F \\
&\leq 3(||\mathcal{A} - \mathcal{A}_{l_1}||_F + 2\sqrt{||\mathcal{N}_{l_1}||_F||\mathcal{A}||_F} + ||\mathcal{N}_{l_1}||_F) \\
&\leq 3(||\mathcal{A} - \mathcal{A}_r||_F + ||\mathcal{A}_r - \mathcal{A}_{l_1}||_F + 2\sqrt{||\mathcal{N}_{l_1}||_F||\mathcal{A}||_F} + ||\mathcal{N}_{l_1}||_F) \\
&\leq 3(||\mathcal{A} - \mathcal{A}_r||_F + \sqrt{\frac{2r}{p}}\kappa||\mathcal{A}||_F + 2\sqrt{||\mathcal{N}_{l_1}||_F||\mathcal{A}||_F} + ||\mathcal{N}_{l_1}||_F) \\
&\leq 3(\epsilon_0 + \underbrace{\sqrt{\frac{2r}{p}}\kappa}_{(\star)} + 2\sqrt{\tilde{\epsilon}} + \tilde{\epsilon})||\mathcal{A}||_F, \qquad (8)
\end{aligned}$$

in the last line we have used the assumption $||\mathcal{N}_{l_1}||_F \leq \tilde{\epsilon}||\mathcal{A}||_F$. In order to choose κ, we use the constraint Eq. 7, fix the sample probability p temporarily, and use the assumption $||\mathcal{N}_{l_1}||_F \leq \sqrt{\frac{l_1\, C_0}{p}} = \tilde{\epsilon}||\mathcal{A}||_F$. It gives

$$l_1 = \frac{p(\tilde{\epsilon}||\mathcal{A}||_F)^2}{C_0} \quad \Rightarrow \quad \kappa^2 \leq \frac{C_0}{p(\tilde{\epsilon}||\mathcal{A}||_F)^2}.$$

Plug the above inequality of κ into the $(\star)$ term of Eq. 8, and requires that $(\star) \leq \epsilon_1$ for some small $\epsilon_1 > 0$, we have

$$||\mathcal{A} - \hat{\mathcal{A}}_{|\cdot|\geq\sigma}||_F \leq 3(\epsilon_0 + \epsilon_1 + 2\sqrt{\tilde{\epsilon}} + \tilde{\epsilon})||\mathcal{A}||_F,$$

where the sample probability must satisfy $p \geq \frac{\sqrt{2rC_0}}{\epsilon_1\tilde{\epsilon}||\mathcal{A}||_F} := p_2$.

Combine two situations we have $||\mathcal{A} - \hat{\mathcal{A}}_{|\cdot|\geq\sigma}||_F \leq \epsilon||\mathcal{A}||_F$, where $\epsilon := 3(\epsilon_0 + \epsilon_1 + 2\sqrt{\tilde{\epsilon}} + \tilde{\epsilon})$, if the sample probability and the threshold are chosen as

$$\begin{aligned}
&p = \max\{p_1, p_2\} \\
&\tau = \kappa||\hat{\mathcal{A}}||_F \leq \sqrt{\frac{C_0}{p\tilde{\epsilon}^2}}\frac{||\hat{\mathcal{A}}||_F}{||\mathcal{A}||_F} \leq \frac{\sqrt{2C_0}}{p\tilde{\epsilon}}.
\end{aligned}$$

■

The above estimation on the error bound in the case of projected tensor SVD is crucial for the quantum algorithm, since quantum projection depends only on the threshold defined for the singular values.

A.3 Data Structure

Theorem A 3. *[8] Let* $\mathbf{x} \in \mathbb{R}^N$ *be a real-valued vector. The quantum state* $|x\rangle = \frac{1}{||\mathbf{x}||_2} \sum_{i=1}^{N} x_i |i\rangle$ *can be prepared using* $\lceil \log N \rceil$ *qubits in time* $\mathcal{O}(\log N)$.

Theorem A 3 claims that there exists a classical memory structure for implementing qRAM. Figure 1 illustrates a simplest example: Given a $N = 4$ dimensional real-valued vector, the quantum state $|x\rangle = x_1 |00\rangle + x_2 |01\rangle + x_3 |10\rangle + x_4 |11\rangle$ can be created by querying the classical structure and applying 3 quantum (controlled) rotations.

Assume that $\mathbf{x}$ is normalized, $||\mathbf{x}||_2 = 1$. The quantum state $|x\rangle$ is created from the initial state $|0\rangle |0\rangle$ by querying the memory structure from the root to the leaf. The first rotation is applied on qubit 1, giving $(\cos\theta_1 |0\rangle + \sin\theta_1 |1\rangle) |0\rangle = (\sqrt{x_1^2 + x_2^2} |0\rangle + \sqrt{x_3^2 + x_4^2} |1\rangle) |0\rangle$, where $\theta_1 := \tan^{-1} \sqrt{\frac{x_3^2 + x_4^2}{x_1^2 + x_2^2}}$. The second rotation is applied on qubit 2 conditioned on the state of qubit 1. It gives

$$\sqrt{x_1^2 + x_2^2} |0\rangle \frac{1}{\sqrt{x_1^2 + x_2^2}} (|x_1| |0\rangle + |x_2| |1\rangle) + \sqrt{x_3^2 + x_4^2} |1\rangle \frac{1}{\sqrt{x_3^2 + x_4^2}} (|x_3| |0\rangle + |x_4| |1\rangle).$$

The last rotation load the signs for the coefficients of conditioned on qubits 1 and 2. In general, a N-dimensional real-valued vector needs to the stored in the classical memory structure with $\lceil \log N \rceil + 1$. Thus, the data vector can be loaded into quantum state using $\mathcal{O}(\lceil \log N \rceil)$ non-trivial controlled rotations.

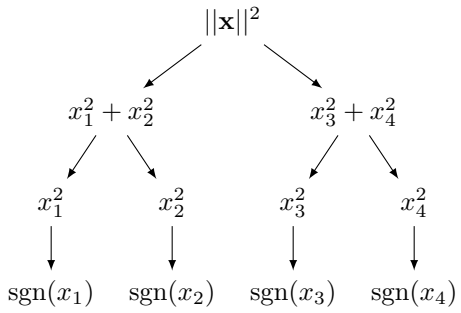

Figure 1: Classical memory structure with quantum access for creating the quantum state $|x\rangle = x_1 |00\rangle + x_2 |01\rangle + x_3 |10\rangle + x_4 |11\rangle$.

Remark: The above simple case of qRAM, generating quantum state from a real-valued vector, can be simply generalized to quantumly accessing matrices or tensors.

A.4 Preparation of $\rho_{\hat{\chi}^\dagger \hat{\chi}}$

Proof. Since the normalized $\hat{\chi} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ is a real-valued tensor, the quantum state $\sum_{i_1 i_2 i_3} \hat{\chi}_{i_1 i_2 i_3} |i_1 i_2 i_3\rangle = \sum_{i_1 i_2 i_3} \hat{\chi}_{i_1 i_2 i_3} |i_1\rangle \otimes |i_2\rangle \otimes |i_3\rangle$ can be prepared via qRAM in time $\mathcal{O}(\log(d_1 d_2 d_3))$. The corresponding density matrix can be written as

$$\rho = \sum_{i_1 i_2 i_3} \sum_{i_1' i_2' i_3'} \hat{\chi}_{i_1 i_2 i_3} |i_1\rangle \otimes |i_2\rangle \otimes |i_3\rangle \langle i_1'| \otimes \langle i_2'| \otimes \langle i_3'| \hat{\chi}_{i_1' i_2' i_3'}.$$

Then, a partial trace on the first index register of the density matrix gives

$$\begin{aligned} \mathrm{tr}_1(\rho) &= \sum_{i_2 i_3} \sum_{i_2' i_3'} \sum_{i_1} \hat{\chi}_{i_1 i_2 i_3} |i_2\rangle \otimes |i_3\rangle \langle i_2'| \otimes \langle i_3'| \hat{\chi}_{i_1 i_2' i_3'} \\ &= \sum_{i_2 i_3 i_2' i_3'} \sum_{i_1} \hat{\chi}_{i_1 i_2 i_3} \hat{\chi}_{i_1 i_2' i_3'} |i_2 i_3\rangle \langle i_2' i_3'| := \rho_{\hat{\chi}^\dagger \hat{\chi}}. \end{aligned}$$

■

A.5 Simulation of the unitary operator $\mathrm{e}^{-it\tilde{\rho}_{\hat{\chi}^\dagger\hat{\chi}}}$

Proof. Recall that $\rho_{\hat{\chi}^\dagger\hat{\chi}} = \sum_{i_2 i_3 i_2' i_3'} \mathcal{C}_{i_2 i_3 i_2' i_3'} |i_2 i_3\rangle \langle i_2' i_3'|$, where $\mathcal{C}_{i_2 i_3 i_2' i_3'} = \sum_{i_1} \hat{\chi}^\dagger_{i_1, i_2 i_3} \hat{\chi}_{i_1, i_2' i_3'}$. For the sake of simplicity, we rewrite $\rho_{\hat{\chi}^\dagger\hat{\chi}}$ as $A \in \mathbb{R}^{N^2 \times N^2}$, where $N := d_2 d_3$. Suppose that the unitary operator needs to be applied on the quantum state $|x\rangle$ whose density matrix reads $\sigma := |x\rangle \langle x|$. Then follow the method in [9], we first create a modified swap operator

$$S_A = \sum_{j,k=1}^{N} A_{jk} |k\rangle \langle j| \otimes |j\rangle \langle k| ,$$

and another auxiliary density matrix $\mu = |\vec{1}\rangle \langle \vec{1}|$, with $|\vec{1}\rangle := \frac{1}{\sqrt{N}} \sum_{k=1}^{N} |k\rangle$. Consider the evolution of the system $\mu \otimes \sigma$ under the unitary operator $\mathrm{e}^{-iS_A \Delta t}$ for a small step Δt. It can be shown that

$$\mathrm{tr}_1\{\mathrm{e}^{-iS_A\Delta t} \mu \otimes \sigma \mathrm{e}^{iS_A \Delta t}\} \approx \mathrm{e}^{-i\frac{A}{N}\Delta t} \sigma \mathrm{e}^{i\frac{A}{N}\Delta t}.$$

Moreover, repeated applications of $\mathrm{e}^{-iS_A\Delta t}$, say n times with $t := n\Delta t$, on the bigger system $\mu \otimes \sigma$ can give $\mathrm{e}^{-i\frac{A}{N}t} \sigma \mathrm{e}^{i\frac{A}{N}t}$ with is the density matrix of the quantum state $\mathrm{e}^{-i\frac{A}{N}t} |x\rangle$. In other words, we can simulate the unitary operator $\mathrm{e}^{-it\tilde{\rho}_{\hat{\chi}^\dagger\hat{\chi}}}$ with $\tilde{\rho}_{\hat{\chi}^\dagger\hat{\chi}} := \frac{\rho_{\hat{\chi}^\dagger\hat{\chi}}}{d_2 d_3}$.

Furthermore, [9] shows that given t and the required accuracy ϵ, the step size Δt should be small enough, such that $n = \mathcal{O}(\frac{t^2}{\epsilon})$. In addition, the quantum access for obtaining the density $\rho_{\hat{\chi}^\dagger\hat{\chi}}$ and creating the modified swap operator requires $T_\rho = \mathcal{O}(\mathrm{polylog}(d_1 d_2 d_3))$ steps. In summary, the total run time for simulating $\mathrm{e}^{-it\tilde{\rho}_{\hat{\chi}^\dagger\hat{\chi}}} |x\rangle$ is $nT_\rho = \mathcal{O}(\frac{t^2}{\epsilon}\mathrm{polylog}(d_1 d_2 d_3))$. ∎

A.6 Quantum Phase Estimation

Theorem A 4 (Phase Estimation [4])**.** *Let unitary $U |v_j\rangle = \mathrm{e}^{i\theta_j} |v_j\rangle$ with $\theta_j \in [-\pi, \pi]$ for $j \in [n]$. There is a quantum algorithm that transforms $\sum_{j\in[n]} \alpha_j |v_j\rangle \mapsto \sum_{j\in[n]} \alpha_j |v_j\rangle |\bar{\theta}_j\rangle$ such that $|\bar{\theta}_j - \theta_j| \leq \epsilon$ for all $j \in [n]$ with probability $1 - 1/\mathrm{poly}(n)$ in time $\mathcal{O}(T_U \log(n)/\epsilon)$, where T_U is the time to implement U.*

A.7 Classical Experiments with Tensor SVD

Some classical results of tensor SVD on different datasets are provided in this section. Given a semantic triple $(\mathrm{s}, \mathrm{p}, \mathrm{o})$, the value function is defined as $\eta_{\mathrm{spo}} = \sum_{i=1}^{R} \sigma_i \, u_{\mathrm{s}}^{(i)} \, u_{\mathrm{p}}^{(i)} \, u_{\mathrm{o}}^{(i)}$, where u_{s}, u_{p}, u_{o} are vector representations of s, p, o, respectively (More details can be found in the review [6]). The model is trained by minimizing the following loss function via stochastic gradient descent,

$$\mathcal{L} = \frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{(\mathrm{s},\mathrm{p},\mathrm{o}) \in \mathcal{D}_{\text{train}}} (y_{\mathrm{spo}} - \eta_{\mathrm{spo}})^{2\alpha} + \gamma(||U_s^\intercal U_s - \mathbb{I}_R||_F + ||U_p^\intercal U_p - \mathbb{I}_R||_F + ||U_o^\intercal U_o - \mathbb{I}_R||_F),$$

where the hyper-parameter γ is used to encourage the orthogonality of embedding matrices, and $\alpha \in \mathbb{Z}$ is a hyper-parameter of the loss function. We compare the tensor SVD model with other benchmark methods, e.g., RESCAL [7], Tucker, and ComplEx [12]. Recall scores of different models are given in Table 1.

	KINSHIP			FB15K-237		
Methods	MR	@3	@10	MR	@3	@10
RESCAL	3.2	88.8	95.5	291.3	20.7	35.1
TUCKER	2.9	89.8	95.0	276.1	20.9	35.7
COMPLEX	2.2	90.0	97.7	242.7	25.2	39.7
TSVD	2.7	84.8	96.6	404.0	21.3	37.1

Table 1: Different recall scores (Mean Rank, Hits@3, Hits@10) of various models on the KINSHIP [2] and FB15K-237 [11] datasets.

A.8 Quantum Algorithm

Algorithm 1 Quantum Tensor SVD for Inference on Knowledge Graph

Input: Inference task $(\mathrm{s}, \mathrm{p}, ?)$
Output: Possible objects to the inference task

Require: Quantum access to $\hat{\chi}$ stored in a classical memory structure; threshold τ for the singular values projection

1: Create $\rho_{\hat{\chi}^\dagger \hat{\chi}}$ via qRAM
2: Create state $|\hat{\chi}_{\mathrm{s}}^{(1)}\rangle_I$ on the input register
3: Prepare unitary operator $U = \sum_{k=0}^{K-1} |k\ \Delta t\rangle \langle k\ \Delta t|_C \exp(-ik\ \Delta t\ \tilde{\rho}_{\hat{\chi}^\dagger \hat{\chi}})$ and apply on $|\hat{\chi}_{\mathrm{s}}^{(1)}\rangle_I$
4: Quantum phase estimation on the clock register to get $\sum_{i=1}^{R} \beta_i |\lambda_i\rangle_C \otimes |u_2^{(i)}\rangle_I \otimes |u_3^{(i)}\rangle_I$
5: Controlled computation on the register to get $\sum_{i=1}^{R} \beta_i |\sigma_i^2\rangle_C \otimes |u_2^{(i)}\rangle_I \otimes |u_3^{(i)}\rangle_I$
6: Singular values projection given the threshold τ to get $\sum_{i:\sigma_i^2 \geq \tau^2} \beta_i |\sigma_i^2\rangle_C \otimes |u_2^{(i)}\rangle_I \otimes |u_3^{(i)}\rangle_I \otimes |0\rangle_R + \sum_{i:\sigma_i^2 < \tau^2} \beta_i |\sigma_i^2\rangle_C \otimes |u_2^{(i)}\rangle_I \otimes |u_3^{(i)}\rangle_I \otimes |1\rangle_R$
7: Uncompute the clock register and measure on the register R and post select the state $|0\rangle_R$
8: Measure the resulting state $\sum_{i:\sigma_i^2 \geq \tau^2} \beta_i |u_2^{(i)}\rangle_I \otimes |u_3^{(i)}\rangle_I = |\hat{\chi}^+_{|\cdot|\geq\tau} \hat{\chi}_{|\cdot|\geq\tau} \hat{\chi}_{\mathrm{s}}^{(1)}\rangle_I$
9: Post-select on p from the sampled triples $(\mathrm{s}, \cdot, \cdot)$

References

[1] Dimitris Achlioptas and Frank McSherry. Fast computation of low-rank matrix approximations. *Journal of the ACM (JACM)*, 54(2):9, 2007.

[2] Arthur Asuncion and David Newman. Uci machine learning repository, 2007.

[3] Iordanis Kerenidis and Anupam Prakash. Quantum recommendation systems. *arXiv preprint arXiv:1603.08675*, 2016.

[4] A Yu Kitaev. Quantum measurements and the abelian stabilizer problem. *arXiv preprint quant-ph/9511026*, 1995.

[5] Nam H Nguyen, Petros Drineas, and Trac D Tran. Tensor sparsification via a bound on the spectral norm of random tensors. *arXiv preprint arXiv:1005.4732*, 2010.

[6] Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33, 2016.

[7] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. A three-way model for collective learning on multi-relational data. In *ICML*, volume 11, pages 809–816, 2011.

[8] Anupam Prakash. *Quantum algorithms for linear algebra and machine learning.* PhD thesis, UC Berkeley, 2014.

[9] Patrick Rebentrost, Adrian Steffens, Iman Marvian, and Seth Lloyd. Quantum singular-value decomposition of nonsparse low-rank matrices. *Physical review A*, 97(1):012327, 2018.

[10] Ryota Tomioka and Taiji Suzuki. Spectral norm of random tensors. *arXiv preprint arXiv:1407.1870*, 2014.

[11] Kristina Toutanova and Danqi Chen. Observed versus latent features for knowledge base and text inference. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*, pages 57–66, 2015.

[12] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In *International Conference on Machine Learning*, pages 2071–2080, 2016.