PAC-Bayesian Generalization Bounds for Knowledge Graph Representation Learning

▼ BDILab ▼ GitHub

Jaejun Lee, **Minsung Hwang,** and **Joyce Jiyoung Whang***
School of Computing, KAIST

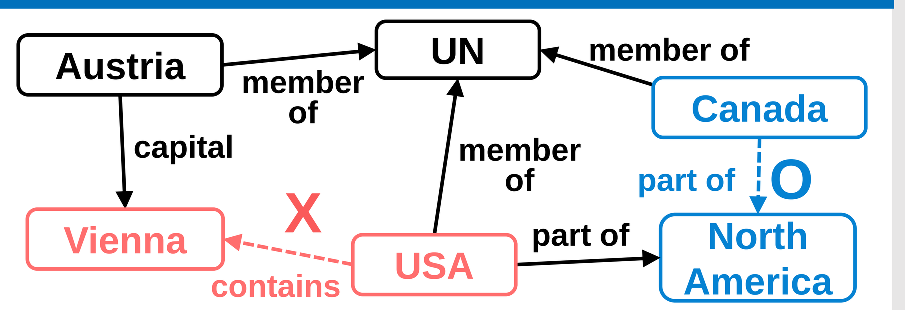
The 41st International Conference on Machine Learning (ICML 2024)

Main Contributions

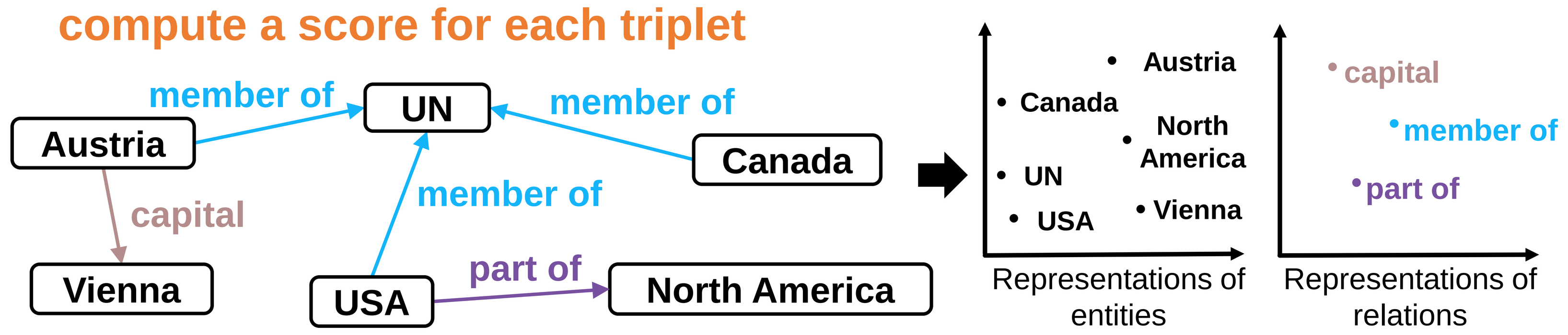
- **ReED framework** representing at least 15 different KGRL methods
  - **RAMP encoder** in ReED is a comprehensive neural encoder for KGRL that can express models such as CompGCN and R-GCN
  - Formulate two types of **triplet classification decoders** in ReED
- Prove the **generalization bounds** for the ReED framework
  - The first study about **PAC-Bayesian generalization bounds** for KGRL
  - **Analyze theoretical findings** from a practical model design perspective
- **Empirically show** that the critical factors in generalization bounds can **explain actual generalization errors** on three real-world KGs

Knowledge Graphs (KGs)

- **Triplet classification on a KG**
  - A model determines whether a given triplet is **plausible or not**

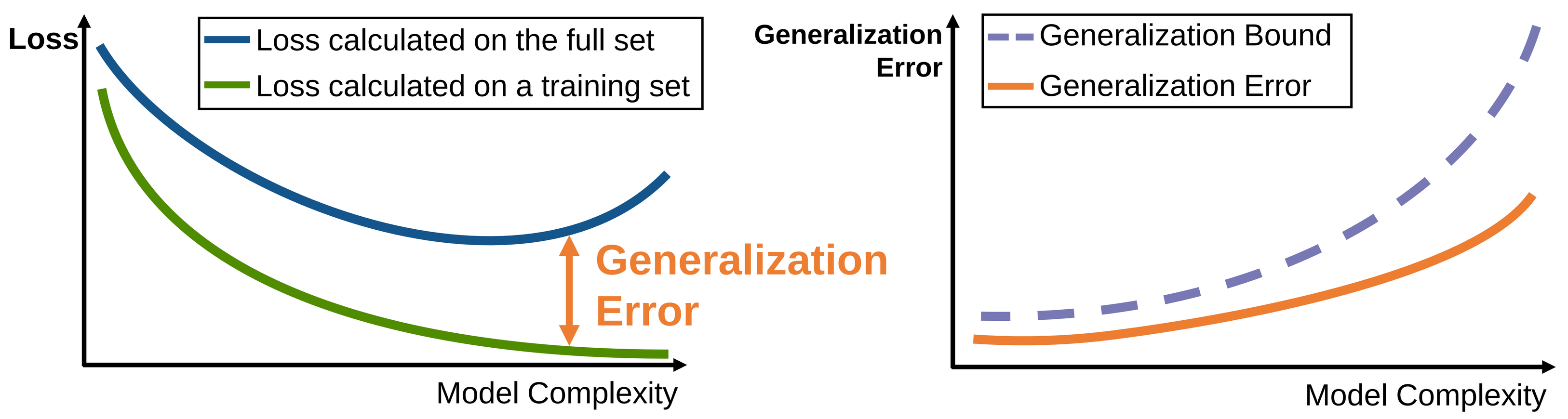


- **Knowledge Graph Representation Learning (KGRL)**
  - **By learning representations** of the entities and relations, KGRL methods **compute a score for each triplet**



Generalization Bound

- **Generalization Error**
  - **Difference** between the losses computed on the **full set** and a **training set**
- **Generalization Bound**
  - **Theoretical upper bound** of the generalization error



Transductive PAC-Bayesian Generalization Bounds

- **Probably Approximately Correct (PAC) Theory**
  - Fundamental tools for analyzing the **generalization bounds**
- **PAC-Bayesian Generalization Bounds**
  - Based on the difference between the **prior and posterior distributions**
- **Transductive PAC-Bayesian Generalization Bounds**
  - Training triplets are **sampled without replacement** from the **finite** full set

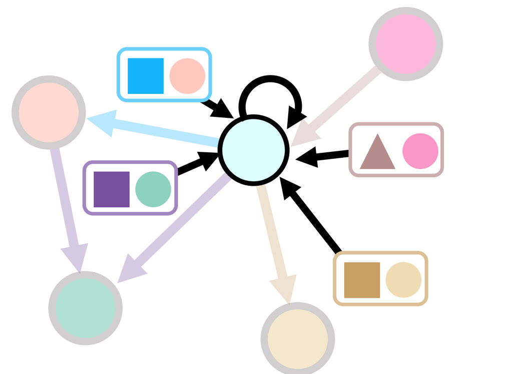
Relation-aware Encoder-Decoder Framework (ReED)

- **R**elation-**A**ware **M**essage **P**assing **Encoder** **(RAMP Encoder)**
  - **Aggregating representations** of the neighboring entities and relations

$$M_r^{(l)}[v,:] = [H^{(l-1)}[v,:] \quad R^{(l-1)}[r,:]] \qquad v \in \mathcal{V}, r \in \mathcal{R}$$

$$H^{(l)} = \phi\left(H^{(l-1)}W_0^{(l)} + \rho\left(\sum_{r\in\mathcal{R}} S_r^{(l)}\psi\left(M_r^{(l)}\begin{bmatrix}W_r^{(l)}\\ U_r^{(l)}\end{bmatrix}\right)\right)\right)$$

$$R^{(l)} = R^{(l-1)}U_0^{(l)}$$



- **Triplet Classification Decoder**
  - Using the entity and relation representations, **compute scores of triplets**
  - **Translational Distance Decoder** (TD decoder)
    - **Distance** between h and t after a relation-specific translation is carried out

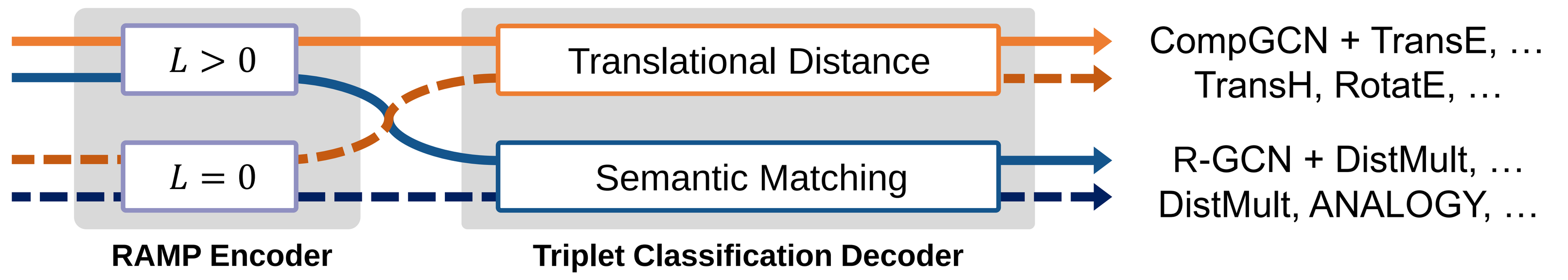
$$f_{\mathbf{w}}(h,r,t)[j] = -\left\| H^{(L)}[h,:]\overline{W}_r^{(j)} + R^{(L)}[r,:]\overline{U}_r^{(j)} - H^{(L)}[t,:]V_r^{(j)} \right\|_2$$

  - **Semantic Matching Decoder** (SM decoder)
    - **Similarity** between the individual components of the triplet

$$f_{\mathbf{w}}(h,r,t)[j] = H^{(L)}[h,:]\overline{U}_r^{(j)}\left(H^{(L)}[t,:]\right)^{\top}$$

Instantiations of ReED

- ReED can express various KGRL methods using **different combinations** of the RAMP encoder and the triplet classification decoder



Empirical and Expected Losses of a Triplet Classifier

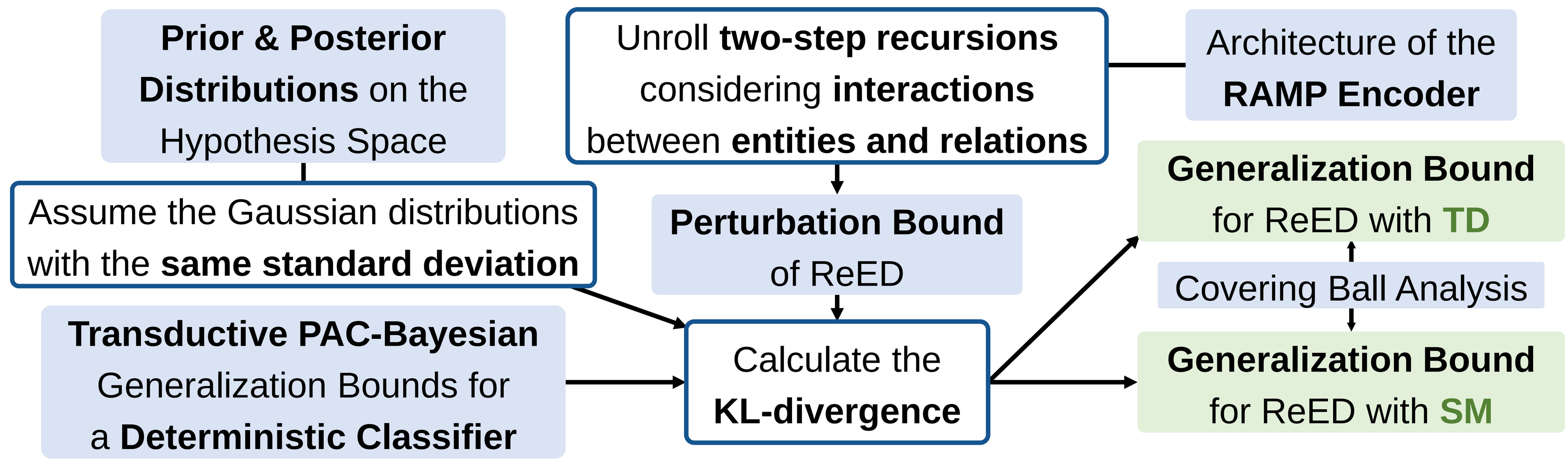
- **Empirical Loss** of Triplet Classifier $f_{\mathbf{w}}$ **:** Measured on a **training triplet set** $\hat{\mathcal{E}}$

$$\mathcal{L}_{\gamma,\hat{\mathcal{E}}}(f_{\mathbf{w}}) = \frac{1}{|\hat{\mathcal{E}}|}\sum_{(h,r,t)\in\hat{\mathcal{E}}} \mathbf{1}\left[f_{\mathbf{w}}(h,r,t)[y_{hrt}] \leq \gamma + f_{\mathbf{w}}(h,r,t)[1-y_{hrt}]\right]$$

- **Expected Loss** of Triplet Classifier $f_{\mathbf{w}}$: Measured on the **full triplet set** $\mathcal{E}$

$$\mathcal{L}_{0,\mathcal{E}}(f_{\mathbf{w}}) = \frac{1}{|\mathcal{E}|}\sum_{(h,r,t)\in\mathcal{E}} \mathbf{1}\left[f_{\mathbf{w}}(h,r,t)[y_{hrt}] \leq f_{\mathbf{w}}(h,r,t)[1-y_{hrt}]\right]$$

Generalization Bounds for ReED: Proof Sketch



PAC-Bayesian Generalization Bounds for ReED

- The generalization bounds for **ReED** with the **TD decoder and SM decoder**

Theorem 4.4 & 4.5 For any $L \geq 0$, let $f_{\mathbf{w}}: \mathcal{V}\times\mathcal{R}\times\mathcal{V} \to \mathbb{R}^2$ be a triplet classifier designed by the combination of the RAMP encoder with L-layers and the triplet classification decoder. Let k_r be the maximum of the infinity norms for all possible $S_r^{(l)}$ in the RAMP encoder. Then, for any $\delta, \gamma > 0$, with probability at least $1-\delta$ over a training triplet set $\hat{\mathcal{E}}$, for any $\mathbf{w}$, we have

$$\mathcal{L}_{0,\mathcal{E}}(f_{\mathbf{w}}) \leq \mathcal{L}_{\gamma,\hat{\mathcal{E}}}(f_{\mathbf{w}}) + \begin{cases} \mathcal{O}\left(\sqrt{\left(\frac{1}{|\hat{\mathcal{E}}|}-\frac{1}{|\mathcal{E}|}\right)\left[\frac{N_{\mathbf{w}}L^2\zeta_L^2 s^{2L} d\ln(N_{\mathbf{w}}d)}{\gamma^2} + \ln\frac{\theta(|\hat{\mathcal{E}}|,|\mathcal{E}|)}{\delta}\right]}\right) & \textbf{(TD)} \\ \mathcal{O}\left(\sqrt{\left(\frac{1}{|\hat{\mathcal{E}}|}-\frac{1}{|\mathcal{E}|}\right)\left[\frac{N_{\mathbf{w}}L^2\eta_L^4 s^{4L} d\ln(N_{\mathbf{w}}d)}{\gamma^2} + \ln\frac{\theta(|\hat{\mathcal{E}}|,|\mathcal{E}|)}{\delta}\right]}\right) & \textbf{(SM)} \end{cases}$$

where $\theta(|\hat{\mathcal{E}}|,|\mathcal{E}|) = 3\sqrt{|\hat{\mathcal{E}}|\left(1-\frac{|\hat{\mathcal{E}}|}{|\mathcal{E}|}\right)}\ln|\hat{\mathcal{E}}|$, $\zeta_L = 2\tau^L\|X_{\mathrm{ent}}\|_2 + 2\kappa\|X_{\mathrm{ent}}\|_2\left(\sum_{i=0}^{L-1}\tau^i\right) + \|X_{\mathrm{rel}}\|_2$, $\eta_L = \tau^L\|X_{\mathrm{ent}}\|_2 + \kappa\|X_{\mathrm{rel}}\|_2\left(\sum_{i=0}^{L-1}\tau^i\right)$, $\tau = C_\phi + \kappa$, $\kappa = C_\phi C_\rho C_\psi \sum_{r\in\mathcal{R}} k_r$, $N_{\mathbf{w}}$ is the total number of learnable matrices, d is the maximum dimension, and s is the maximum Frobenius norm of the learnable matrices

Generalization Bounds for ReED: a Simplified Form

- Leaving **model-dependent terms** and regarding the rest as a constant

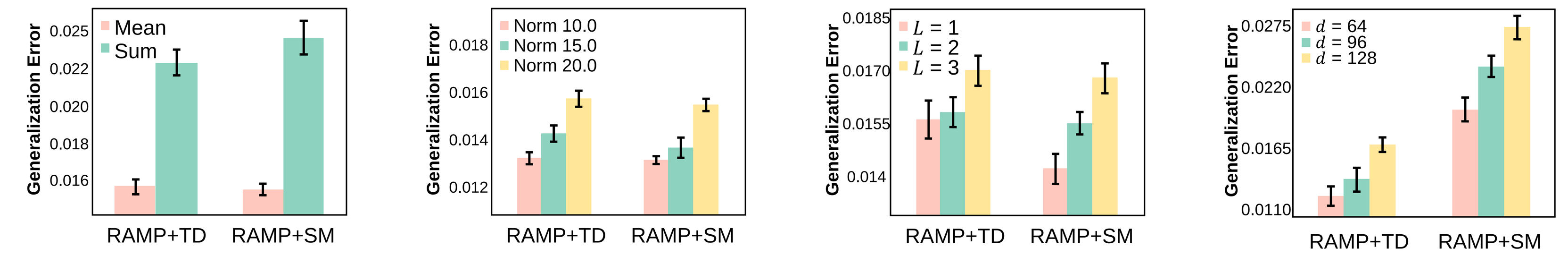
$$\mathcal{L}_{0,\mathcal{E}}(f_{\mathbf{w}}) \leq \mathcal{L}_{\gamma,\hat{\mathcal{E}}}(f_{\mathbf{w}}) + \begin{cases} \mathcal{O}\left(L\left(\sum_{r\in\mathcal{R}}k_r\right)^L s^L\sqrt{N_{\mathbf{w}}\ln N_{\mathbf{w}}}\right) & \textbf{(TD)} \\ \mathcal{O}\left(L\left(\sum_{r\in\mathcal{R}}k_r\right)^{2L} s^{2L}\sqrt{N_{\mathbf{w}}\ln N_{\mathbf{w}}}\right) & \textbf{(SM)} \end{cases}$$

- **Practical implications** that can guide the desirable designs of KGRL
  - k_r: **Maximum of the infinity norms** for all possible $S_r^{(l)}$
    - A **mean aggregator** can be a better option than a **sum aggregator**
  - $N_{\mathbf{w}}$: **Total number** of learnable matrices ($= \mathcal{O}(|\mathcal{R}|L)$)
    - **Parameter-sharing** strategies & **basis/block decomposition** ideas
  - s: **Maximum Frobenius norm** of the learnable matrices
    - **Weight normalization** & **Normalization of entity representations**

Experimental Results

- **Measure the generalization errors** on real-world knowledge graphs

Varying **the aggregator**: Mean vs Sum | Varying **the norm of weight matrices** | Varying **the number of layers** | Varying **the maximum dimension**



→ The **critical factors** also affect an **actual generalization error**

Conclusion

- A novel **ReED framework** expressing at least 15 KGRL methods
- The **first PAC-Bayesian generalization bounds** for ReED with two different types of decoders: **TD decoder** and **SM decoder**
- Provide **theoretical grounds** for **commonly used tricks** in KGRL
- Empirically show the relationship between **the critical factors in the theoretical bounds** and **the actual generalization errors**