# Supplementary Material to "Augmenting and Tuning Knowledge Graph Embeddings"

**Robert Bamler***
Department of Computer Science
University of California, Irvine
Irvine, CA 92617

**Farnood Salehi***
École Polytechnique Fédérale
de Lausanne (EPFL),
Switzerland

**Stephan Mandt**
Department of Computer Science
University of California, Irvine
Irvine, CA 9261

## The Role of Parameter Uncertainty in the Proposed Hyperparameter Optimization

Bayesian inference and the idea of measuring uncertainty are somewhat uncommon in the literature for knowledge graph embeddings. To clarify why uncertainty is important in the proposed hyperparameter optimization (Section 3.1 of the main text), we compare the method here to a more naive approach that will turn out to fail because it ignores uncertainty. We discuss the failure of the naive approach and the benefit of estimating parameter uncertainty first intuitively and then more formally.

**Intuitive picture.** The variational EM algorithm maximizes (a lower bound on) the marginal likelihood $p(\mathbb{S}'|\boldsymbol{\lambda})$, Eq. 14 of the main text. In a more naive attempt to hyperparameter tuning without cross validation, one might be tempted to skip the marginalization over model parameters $\mathbf{E}$ and $\mathbf{R}$. Instead, one might try to directly maximize the log joint probability $\log p(\mathbf{E}, \mathbf{R}, \mathbb{S}'|\boldsymbol{\lambda})$, i.e., minimize the loss $L = -\log p(\mathbf{E}, \mathbf{R}, \mathbb{S}'|\boldsymbol{\lambda})$, over $\mathbf{E}$, $\mathbf{R}$, and hyperparameters $\boldsymbol{\lambda}$. This approach, however, would lead to divergent solutions because the log joint probability is unbounded.

The log joint probability contains the log priors (first two terms on the right-hand side of Eq. 10 of the main text). Figure S1 shows the prior $p(\mathrm{E}_{ek}|\lambda^{\mathrm{E}}_e)$, Eq. 8 of the main text, of a single component $\mathrm{E}_{ek}$ of an entity embedding assuming, for simplicity, a real embedding space. With growing regularizer strength (increasing $\lambda^{\mathrm{E}}_e$), the prior becomes narrower and narrower. As the peak narrows, it also grows higher due to the normalization constraint. In the limit $\lambda^{\mathrm{E}}_e \to \infty$, the prior collapses to an infinitely narrow and high $\delta$-peak at zero.

A hyperparameter tuning method that ignores posterior uncertainty can exploit the unbounded growth of the maximum of the prior to send the log joint distribution to infinity (i.e., the loss $L \to -\infty$). Without any posterior uncertainty, one can set the model parameter $\mathrm{E}_{ek}$ precisely to zero and then make the value of $p(\mathrm{E}_{ek}|\lambda^{\mathrm{E}}_e)$ arbitrarily large by sending $\lambda^{\mathrm{E}}_{ek} \to \infty$.

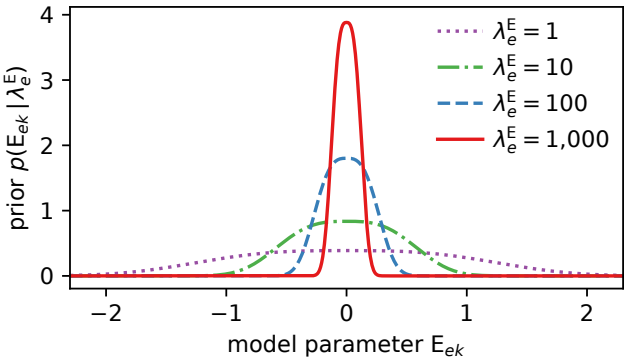


Figure S1: Prior (Eq. 8 of the main text with $p = 3$) for a single model parameter $\mathrm{E}_{ek}$. As the prior gets more peaked for growing regularizer strength $\lambda^{\mathrm{E}}_e$, the height of the peak grows unboundedly. An (incorrect) hyperparameter optimization method that ignores parameter uncertainty could end up exploiting this unboundedness and diverge to $\lambda^{\mathrm{E}}_e \to \infty$.

We prevent this collapse of the prior to a $\delta$-peak by keeping track of parameter uncertainty. Admitting a nonzero uncertainty for $\mathrm{E}_{ek}$ no longer allows us to set $\mathrm{E}_{ek}$ precisely and deterministically to zero. Any slightly nonzero value of $\mathrm{E}_{ek}$ would have no support under a $\delta$-peaked prior.

**Formal derivation.** We now formalize the above intuitive picture and show that the specific variational approximation chosen in Eqs. 12-13 indeed suffices to prevent any divergent solutions.

Assuming again a real embedding space, the log prior of a single model parameter $\mathrm{E}_{ek}$ is given by (cf., Eq. 16 of the main text)

$$\log p(\mathrm{E}_e|\lambda^{\mathrm{E}}_e) = \frac{1}{p}\left[\log \lambda^{\mathrm{E}}_e - \lambda^{\mathrm{E}}_e|\mathrm{E}_{ek}|^p\right] + \text{const.} \qquad \text{(S1)}$$

As discussed above, setting $\mathrm{E}_{ek} = 0$ and sending $\lambda^{\mathrm{E}}_e \to \infty$ sends the right-hand side of Eq. S1 to infinity. This can

*joint first authorship.

even be relaxed: the log prior diverges for $\lambda_e^{\mathrm{E}} \to \infty$ as long as we keep $\mathrm{E}_{ek}$ small enough, i.e., as long as $\mathrm{E}_{ek} = O((\lambda_e^{\mathrm{E}})^{-1/p})$. This is why maximizing the log joint distribution over $\boldsymbol{\lambda}$ leads to divergend solutions.

Instead of maximizing the log joint distribution, the variational EM algorithm maximizes the ELBO, Eq. 14 of the main text. Note first that the ELBO itself is bounded from above by zero: the ELBO is a lower bound on the marginal log likelihood $\log p(\mathbb{S}'|\boldsymbol{\lambda})$, where $p(\mathbb{S}'|\boldsymbol{\lambda}) \leq 1$ since it is a discrete probability distribution.

Further, the ELBO has a maximum at finite values for the variational parameters and hyperparameters. Maximizing the ELBO ensures that each model parameter is associated with a nonzero uncertainty $\sigma_{e/r}^{\mathrm{E/R}} > 0$ since the entropy term

$$H[q_{\boldsymbol{\mu},\boldsymbol{\sigma}}] = \sum_{e\in[N_{\mathrm{e}}]} \log \sigma_e^{\mathrm{E}} + \sum_{r\in[N_{\mathrm{r}}]} \log \sigma_r^{\mathrm{R}} + \text{const.} \quad \text{(S2)}$$

imposes an infinite penalty if any $\sigma_{e/r}^{\mathrm{E/R}} \to 0$. The entropy term in the ELBO thus has an additional regularizing effect. Combined with the other regularizing term in the ELBO, the expected log prior, we obtain for a given model parameter $\mathrm{E}_{ek}$ using Eq. S1, up to an additive constant,

$$\begin{aligned}
&\mathbb{E}_{q_{\boldsymbol{\mu},\boldsymbol{\sigma}}}\left[\log p(\mathrm{E}_{ek}|\lambda_e^{\mathrm{E}}) - \log\left(q_{\sigma_{ek}^{\mathrm{E}},\mu_{ek}^{\mathrm{E}}}(\mathrm{E}_{ek})\right)\right] \\
&= \frac{1}{p}\log\lambda_e^{\mathrm{E}} - \frac{\lambda_e^{\mathrm{E}}}{p}\mathbb{E}_{q_{\boldsymbol{\sigma},\boldsymbol{\mu}}}\left[|\mathrm{E}_{ek}|^p\right] + \log\sigma_{ek}^{\mathrm{E}} \quad \text{(S3)} \\
&= \frac{1}{p}\left[\log\left(\lambda_e^{\mathrm{E}}\left(\sigma_{ek}^{\mathrm{E}}\right)^p\right) - \lambda_e^{\mathrm{E}}\,\mathbb{E}_{\epsilon\sim\mathcal{N}(0,1)}\left[\left|\mu_{ek}^{\mathrm{E}} + \sigma_{ek}^{\mathrm{E}}\epsilon\right|^p\right]\right]
\end{aligned}$$

where, in the last equality, we made the dependency on $\sigma_{ek}^{\mathrm{E}}$ explict by reparameterizing the normal distributed random variable $\mathrm{E}_{ek} = \mu_{ek}^{\mathrm{E}} + \sigma_{ek}^{\mathrm{E}}\epsilon$ in terms of a standard-normal distributed variable $\epsilon$.

Maximizing the right-hand side of Eq. S3 over $\mu_{ek}^{\mathrm{E}}$ yields $\mu_{ek}^{\mathrm{E}} = 0$ by symmetry, thus simplifying the objective to

$$\begin{aligned}
&\frac{1}{p}\left[\log\left(\lambda_e^{\mathrm{E}}\left(\sigma_{ek}^{\mathrm{E}}\right)^p\right) - \lambda_e^{\mathrm{E}}\left(\sigma_{ek}^{\mathrm{E}}\right)^p\,\mathbb{E}_{\epsilon\sim\mathcal{N}(0,1)}\left[|\epsilon|^p\right]\right] \\
=&\frac{1}{p}\left[\log\left(\lambda_e^{\mathrm{E}}\left(\sigma_{ek}^{\mathrm{E}}\right)^p\right) - \lambda_e^{\mathrm{E}}\left(\sigma_{ek}^{\mathrm{E}}\right)^p c_p\right] \quad \text{(S4)}
\end{aligned}$$

with some numerical constant $c_p > 0$. The right-hand side of Eq. S4 is structurally similar to the right-hand side of Eq. S1, which had divergent solutions: both are a difference between a logarithmic and a linear term in $\lambda_e^{\mathrm{E}}$. However, while in Eq. S1, we were able to send the logarithmic term to infinity and still keep the linear term bounded, this is not possible in Eq. S4, which has the same argument (up to a constant $c_p$) for the logarithmic and the linear term. The right-hand side of Eq. S4 has a maximum at a finite value for $\lambda_e^{\mathrm{E}}\left(\sigma_{ek}^{\mathrm{E}}\right)^p$. Thus, the variational EM algorithm avoids divergent solutions.