Empirical Risk Minimization and Stochastic Gradient Descent for Relational Data

Vicor Veitch[1] **Morgane Austern**[1] **Wenda Zhou**[1] **David M. Blei** **Peter Orbanz**
Columbia University

Abstract

Empirical risk minimization is the main tool for prediction problems, but its extension to relational data remains unsolved. We solve this problem using recent ideas from graph sampling theory to (i) define an empirical risk for relational data and (ii) obtain stochastic gradients for this empirical risk that are automatically unbiased. This is achieved by considering the method by which data is sampled from a graph as an explicit component of model design. By integrating fast implementations of graph sampling schemes with standard automatic differentiation tools, we provide an efficient turnkey solver for the risk minimization problem. We establish basic theoretical properties of the procedure. Finally, we demonstrate relational ERM with application to two non-standard problems: one-stage training for semi-supervised node classification, and learning embedding vectors for vertex attributes. Experiments confirm that the turnkey inference procedure is effective in practice, and that the sampling scheme used for model specification has a strong effect on model performance.

1 Introduction

Relational data is data that can be represented as a graph, possibly annotated with additional information. An example is the link graph of a social network, annotated by user profiles. We consider prediction problems for such data. For example, how to predict the preferences of a user of a social network using both the preferences and profiles of other users, and the network itself? In the classical case of i.i.d. data—where the observed data does not include link structure—the data decomposes into individual examples. Prediction methods for i.i.d. data typically rely on this decomposition, e.g., predicting a user's preferences from only the profile of the user, ignoring the network structure. Relational data, however, does not decompose; e.g., because of the link structure, a social network can not be decomposed into individual users. Accordingly, classical methods do not generally apply to relational data, and new methods cannot be developed with the same ease as for i.i.d. data.

With i.i.d. data, prediction problems are typically solved with models fit by empirical risk minimization (ERM) [21, 22, 19]. We give an (unusual) presentation of ERM that anticipates the relational case. The observed data is a set $\overline{\mathbb{S}}_n = \{\overline{X}_1, \dots, \overline{X}_n\}$ that decomposes into examples $\overline{X}_i = (X_i, Y_i)$. The task is to choose a predictor π that completes X by estimating missing information Y, e.g., a class label. An ERM model is defined by two parts: (i) a hypothesis class $\{\pi_\theta | \theta \in \mathcal{T}\}$ from which π is chosen, and (ii) a loss function L where $L(\bar{x}; \theta) \in \mathbb{R}_+$ measures the reconstruction error of predictor π_θ on example $\bar{x}$. The empirical risk is the expected loss on an example randomly selected from the dataset:

$$\hat{R}(\theta, \overline{\mathbb{S}}_n) \coloneqq \mathbb{E}_{\overline{X} \sim \mathbb{F}(\overline{\mathbb{S}}_n)}[L(\overline{X}; \theta) | \overline{\mathbb{S}}_n], \tag{1}$$

where $\mathbb{F}(\overline{\mathbb{S}}_n)$ is the empirical distribution.[2] The ERM dogma is to select the predictor $\pi_{\hat{\theta}_n}$ given by $\hat{\theta}_n = \operatorname{argmin}_\theta \hat{R}(\theta, \overline{\mathbb{S}}_n)$. That is, the objective function that defines learning is the empirical risk.

ERM has two useful properties. (1) It provides a principled framework for defining new machine learning methods. In particular, when examples are generated i.i.d., model-agnostic results guarantee that ERM models cohere as more data is collected (e.g., in the sense of statistical convergence) [19]. (2) For differentiable

[1] Equal contribution

Preliminary work. Under review by AISTATS 2019. Do not distribute.

[2] The empirical risk is more often equivalently written as $\hat{R}(\theta, \overline{\mathbb{S}}_n) = \frac{1}{n} \sum_{i \le n} L(\overline{X}_i; \theta)$.

models, mini-batch stochastic gradient descent (SGD) can efficiently solve the minimization problem (albeit, approximately). The ease of SGD comes from the definition of the empirical risk as the expectation over a randomly subsampled example: the gradient of the loss on a randomly subsampled example is an unbiased estimate of the gradient of the empirical risk. Combined with automatic differentiation, this provides a turnkey approach to fitting machine-learning models.

Returning to relational data, the observed data is now a graph $\overline{G}_n$ of size n (e.g., the number of vertices or edges). The graph is possibly annotated, e.g., by vertex labels. We further consider G_n as an incomplete version of $\overline{G}_n$. For example, G_n may censor labels of the vertices or some of the edges from $\overline{G}_n$. In relational learning, the task is to find a predictor π that completes G_n by estimating the missing information. Typically, π is chosen from a parameterized family $\{\pi_\theta | \theta \in \mathcal{T}\}$ to minimize an objective function $\mathcal{O}_n(\theta, \overline{G}_n)$. Unlike the empirical risk, the objective $\mathcal{O}_n$ is not built from a loss on individual examples; $\mathcal{O}_n$ must be specified for the entire observed graph.

In relational learning, there is not yet a framework that has properties (1) and (2) of ERM. The challenge is that relational data does not decompose into individual examples. Regarding (1), theory is elusive because the i.i.d. assumption is meaningless for relational data. This makes it difficult to reason about what happens as more data is collected. Regarding (2), mini-batch SGD is not generally applicable even for differentiable models. SGD requires unbiased estimates of the full gradient. For a random subgraph G_k of G_n, the stochastic gradient $\nabla_\theta \mathcal{O}_k(\pi_\theta(G_k), \overline{G}_k)$ is not generally unbiased. In particular, the bias depends on the choice of random sampling scheme used to select the subgraph. Circumventing these two issues requires either careful design of the objective function used for learning [e.g., 16, 7, 3, 25, 8], or model-specific derivation and analysis. For example, graph convolutional networks [10, 11, 18, 20] use full batch gradients, and scaling training requires custom derivation of stochastic gradients [4].

This paper introduces relational ERM, a generalization of ERM to relational data. Relational ERM provides a recipe for machine learning with relational data that preserves the two important properties of ERM:

1. It provides a simple way to define (task-specific) relational learning methods, and
2. For differentiable models, relational ERM minimization can be efficiently solved in a turnkey fashion by mini-batch stochastic gradient descent.

Relational ERM mitigates the need for model-specific analysis and fitting procedures.

Extending turnkey mini-batch SGD to relational data allows the easy use of autodiff-based machine-learning frameworks for relational learning. To facilitate this, we provide fast implementations of a number of graph subsampling algorithms, and integration with TensorFlow.[3]

In Section 2 we define relational ERM models and show how to automatically calculate unbiased mini-batch stochastic gradients. In Section 3 we explain connections to previous work on machine learning for graph data and we illustrate how to develop task-specific relational ERM models. In Section 4 we review several randomized algorithms for subsampling graphs. Relational ERM models require the specification of such algorithms. In Section 5 we establish theory for relational ERM models. The main insights are: (i) the i.i.d. assumption can be replaced by an assumption on how the data is collected [15, 24, 1, 5], and, (ii) the choice of randomized sampling algorithm is necessarily viewed as a model component. In Section 6, we study relational ERM empirically by implementing the models of Section 3. We observe that the turnkey mini-batch SGD procedure succeeds in efficiently fitting the models, and that the choice of graph subsampling algorithm has a large effect in practice.

2 Relational ERM and SGD

Our aim is to define relational ERM in analogy with classical ERM. The fundamental challenge is that relational data does not decompose into individual examples. Classical ERM uses the empirical distribution to define the objective function Eq. (1). There is no canonical analogue of the empirical distribution for relational data.

The first insight is that the empirical distribution may be viewed as a randomized algorithm for subsampling the dataset. The required analogue is then a randomized algorithm for subsampling a graph. In the i.i.d. setting, uniform subsampling is almost always used. However, there are many possible ways to sample from a graph. We review a number of possibilities in Section 4. For example, the sampling algorithm might draw a subgraph induced by sampling k vertices at random, or the subgraph induced by a random walk of length k. The challenge is that there is no a priori criterion for deciding which sampling algorithm is "best."

Our approach is to give up and declare victory: we *define* the required analogue as a *component of model design*. We require the analyst to choose a randomized sampling algorithm Sample, where $\mathsf{Sample}(\overline{G}_n, k)$ is a

[3] Supplementary material.

random subgraph of size k. The choice of Sample defines a notion of "example." This allows us to complete the analogy to classical ERM.

A *relational ERM model* is defined by three ingredients:

1. A sampling routine Sample.
2. A predictor class $\{\pi_\theta | \theta \in \mathcal{T}\}$ with parameter θ.
3. A loss function L, where $L(\overline{G}_k; \theta)$ measures the reconstruction quality of π_θ on example G_k.

The objective function is defined in analogy with the empirical risk Eq. (1). The *relational empirical risk* is:

$$\hat{R}_k(\pi, \overline{G}_n) := \mathbb{E}_{\overline{G}_k = \mathsf{Sample}(\overline{G}_n, k)}[L(\overline{G}_k; \theta) \mid \overline{G}_n]. \quad (2)$$

Relational empirical risk minimization selects a predictor $\hat{\pi}$ that minimizes the relational empirical risk,

$$\hat{\pi} := \pi_{\hat{\theta}_n} \quad \text{where} \quad \hat{\theta}_n := \operatorname*{argmin}_\theta \hat{R}_k(\pi_\theta, \overline{G}_n) \,. \quad (3)$$

Stochastic gradient descent

A crucial property of relational ERM is that SGD can be applied to solve the minimization problem Eq. (3) without any model specific analysis. Define a stochastic gradient as $\nabla_\theta L(\mathsf{Sample}(G_n, k); \theta)$, the gradient of the loss computed on a sample of size k drawn with Sample. Observe that

$$\begin{aligned} \nabla_\theta \hat{R}_r(\theta, G_n) &= \nabla_\theta \mathbb{E}[L(\mathsf{Sample}(G_n, k); \theta) \mid \overline{G}_n] \\ &= \mathbb{E}[\nabla_\theta L(\mathsf{Sample}(G_n, k); \theta) \mid \overline{G}_n]. \end{aligned}$$

That is, the random gradient $\nabla_\theta L(\mathsf{Sample}(G_n, k); \theta)$ is an unbiased estimator of the gradient of the full relational empirical risk. If Sample is computationally efficient, then SGD with this stochastic estimator can solve the relational ERM.

To specify a relational ERM model in practice, the practitioner implements the three ingredients in code. Machine-learning frameworks provide tools to make it easy to specify a class of predictors and a per-example loss function, which are ingredients of classical ERM. Relational ERM additionally requires implementing Sample and integrating it with a machine-learning framework. In practice, Sample can be chosen from a standard library of sampling routines. To that end, we provide efficient implementations of a number of routines and integration with an automatic differentiation framework (TensorFlow).[4] This gives an effective "plug-and-play" approach for defining and fitting models.

[4]Supplementary material.

3 Example Models

We consider several examples of relational ERM models. We split the parameter into a pair $\theta = (\gamma, \lambda)$: the global parameters γ are shared across the entire graph, and the embedding parameters λ provide low-dimensional embeddings λ_v for each vertex v. Informally, global parameters encode population properties—"people with different political affiliation are less likely to be friends"—and the embeddings encode per-vertex information—"Bob is a radical vegan."

Graph representation learning

Methods for learning embeddings of vertices are widely studied; see [9] for a review. Many such methods rely on decomposing the graph into neighborhoods determined by (random) walks of fixed size. One example is Node2Vec [7] (an extension of DeepWalk [16]). The basic approach is to draw a large collection of simple random walks, view each of these walks as a "sentence" where each vertex is a "word", and learn vertex embeddings by applying a standard word embedding method [14, 13]. To use mini-batch SGD, the objective function is restricted to a uniform sum over all walks. Unbiased stochastic gradients to be computed by uniformly sampling walks.

Relational ERM models include graph representation models of this kind. For example, Node2Vec [7] is equivalent to a relational ERM model that (i) predicts graph structure using a predictor parameterized only by embedding vectors, (ii) uses a cross entropy loss on graph structure, and (iii) takes Sample as a random-walk of fixed length (augmented with randomly sampled negative examples).

A number of other relational learning methods also enable SGD by restricting the objective function to a uniform sum over fixed-size subgraphs [e.g., 7, 3, 25, 8]. Any such model is equivalent to a relational ERM model that takes Sample as the uniform distribution over fixed-size subgraphs. But, in general, relational ERM does not require restricting to sampling schemes of this kind. Note that "negative-sampling" algorithms—which are critical in practice—do not uniformly sample fixed size subgraphs.

The next examples illustrate relational ERM for problems that are difficult with existing approaches to graph representation learning.

Semi-supervised node classification

Consider a network G_n where each node i is labeled by binary features—for example, hyperlinked documents labeled by subjects, or interacting proteins labeled by function. The task is to predict the labels of a subset

of these nodes using the graph structure and the labels of the remaining nodes.

The model has the following form: Each vertex i is assigned a k-dimensional embedding vector $\lambda_i \in \mathbb{R}^k$. Labels are predicted using a parameterized function $f(\bullet;\gamma) : \mathbb{R}^k \to [0,1]^L$ that maps the node embeddings to the probability of each label. The presence or absence of edge i,j is predicted based on $\lambda_i^T \lambda_j$. This enables learning embeddings for unlabeled vertices. Let σ denote the sigmoid function; let label $l_{ij} \in \{0,1\}$ denote whether vertex i has label j; and let $q \in [0,1]$. The loss on subgraphs $G_k \subset G_n$ is:

$$L(G_k;\lambda,\gamma,l) = \tag{4}$$
$$q\Big(\sum_{i\in v(G_k)}\sum_{j=1}^{L} l_{ij}\log f(\lambda_i;\gamma)_j + (1-l_{ij})\log(1-f(\lambda_i;\gamma)_j)\Big)$$
$$+(1-q)\Big(-\sum_{i,j\in e(G_k)}\log\sigma(\lambda_j^T\lambda_i) - \sum_{i,j\in\bar{e}(G_k)}\log(1-\sigma(\lambda_j^T\lambda_i))\Big).$$

Here, v, e, and $\bar{e}$ denote the vertices, edges, and non-edges of the graph respectively. The loss on edge terms is cross-entropy, a standard choice in embedding models [9]. Intuitively, the predictor uses the embeddings to predict both the vertex labels and the subgraph structure.

The model is completed by choosing a sampling scheme Sample. Relational ERM then fits the parameters as

$$(\hat{\lambda}_n, \hat{\gamma}_n) = \operatorname*{argmin}_{\lambda,\gamma} \mathbb{E}[L\big(\lambda,\gamma;\ \mathsf{Sample}(G_n,k), l\big) \mid G_n].$$

We can vary the choice of Sample independent of optimization concerns; in Section 6 we observe that this leads to improved predictive performance.

Older embedding approaches use a two-stage procedure: node embeddings are first pre-trained using the graph structure, and then used as inputs to a logistic regression that predicts the labels [e.g., 16, 7]. Yang, Cohen, and Salakhudinov [25] adapt a random-walk based method to allow simultaneous training; their approach requires extensive development, including a custom (two-stage) variant of SGD. Relational ERM allows simultaneous learning with no need for model specific derivation.

Wikipedia category embeddings

We consider Wikipedia articles joined by hyperlinks. Each article is tagged as a member of one or more categories—for example, "`Muscles_of_the_head_and_neck`", "`Japanese_rock_music_groups`", or "`People_from_Worcester`." The task is to learn embeddings that encode semantic relationships between the categories.

Let G_n denote the hyperlink graph and let $\mathcal{C}(i)$ denote the categories of article i. Each category $c \in C$ is assigned an embedding γ_c, and the embedding of each article (vertex) is taken to be the sum of the embeddings of its categories, $\lambda_i \coloneqq \sum_{c\in\mathcal{C}(i)} \gamma_c$. The loss is

$$L(G_k, C; \lambda) = \tag{5}$$
$$-\sum_{i,j\in e(G_k)} \log\sigma(\lambda_j^T\lambda_i) - \sum_{i,j\in\bar{e}(G_k)} \log(1-\sigma(\lambda_j^T\lambda_i)),$$

where e and $\bar{e}$ denote, respectively, the presence and absence of hyperlinks between articles. Intuitively, the predictor uses the category embeddings to predict the hyperlink structure of subgraphs. Relational ERM chooses the embeddings as

$$\hat{\gamma}_n = \operatorname*{argmin}_{\gamma} \mathbb{E}[L\big(\lambda(\gamma);\ \mathsf{Sample}(G_n,k), C\big) \mid G_n].$$

We write $\lambda(\gamma)$ to emphasize that the article embeddings are a function of the category embeddings. Category embeddings obtained with this model are illustrated in Fig. 1; see Section 6 for details on the experiment.

The point of this example is: relational ERM makes it easy to implement this non-standard relational learning model and fit it with mini-batch SGD. The use of mini-batch SGD is important because the data graph is large.

4 Subsampling algorithms

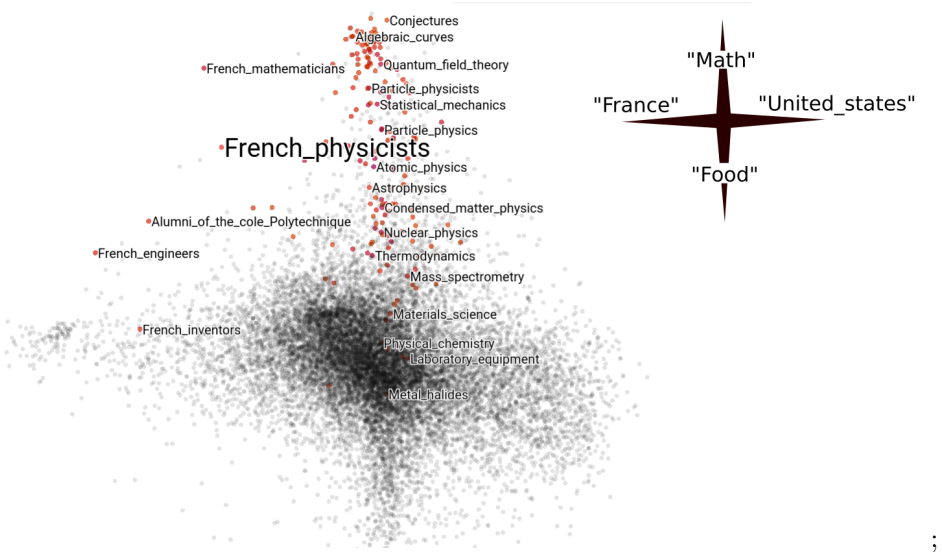
In classical ERM, sampling uniformly (with or without replacement) is typically the only choice. In contrast, there are many ways to sample from a graph. Each such sampling algorithm Sample leads to a different notion of empirical risk in (2).

As described above, random walks underlie graph representation methods built in analogy with language models. A simple random walk of length k on a graph $\overline{G}_n$ selects vertices $v_1,\ldots,v_k$ by starting at a given vertex v_1, and drawing each vertex v_{i+1} uniformly from the neighbors of v_i. Typically, random-walk based methods augment the sample by hallucinating additional edges using a strategy borrowed from the Skipgram model [14]:

Algorithm 1 (Random walk: Skipgram [16])**.**

(i) Sample a random walk $v_1,\ldots,v_k$ starting at a uniformly selected vertex of $\overline{G}_n$.

(ii) Report $\overline{G}_k = \{(v_i,v_j) : d(v_i,v_j) < W\}$. The *window* W is a sampler parameter, and $d(v_i,v_j)$ is the number of steps between v_i and v_j.



;

Figure 1: Trained Wikipedia category embeddings. Category embeddings are projected into a 2-dimensional space, with a projection chosen to maximally separate "`France`" and "`United_states`" horizontally, and "`Math`" and "`Food`" vertically. Highlighted categories are nearest neighbors of "`French_physicists`."

Since relational ERM is indifferent to the connection with language models, a natural alternative augmentation strategy is:

Algorithm 2 (Random walk: Induced)**.**

(i) Sample a random walk $v_1, \ldots, v_k$ starting at a uniformly selected vertex of $\overline{G}_n$.

(ii) Report $\overline{G}_k$ as the edge list of the vertex induced subgraph of the walk.

A simple choice is to sample k vertices uniformly at random and report $\overline{G}_k$ as the induced subgraph. Such an algorithm will not work well in practice since it is not suitable for sparse graphs. We are typically interested in the case $k \ll n$. If $\overline{G}_n$ is sparse then such a sample typically includes few or no edges, and thus carries little information about $\overline{G}_n$. The next algorithm modifies uniform vertex sampling to fix this pathology. The idea is to over-sample vertices and retain only those vertices that participate in at least one edge in the induced subgraph.

Algorithm 3 (p-sampling [24])**.**

(i) Select each vertex in $\overline{G}_n$ independently, with a fixed probability $p \in [0, 1]$.

(ii) Extract the induced subgraph $\overline{G}_k$ of $\overline{G}_n$ on the selected vertices.

(iii) Delete all isolated vertices from $\overline{G}_k$, and report the resulting graph.

Algorithm 3 originates in the foundations of statistical network modeling [24], and has not previously been used in the context of graph representation learning.

Another natural sampling scheme is:

Algorithm 4 (Uniform edge sampling)**.**

(i) Select k edges in $\overline{G}_n$ uniformly and independently from the edge set.

(ii) Report the graph $\overline{G}_k$ consisting of these edges, and all vertices incident to these edges.

4.1 Negative sampling

For a pair of vertices in an input graph $\overline{G}_n$, a sampling algorithm can report three types of edge information: The edge may be observed as present, observed as absent (a *non-edge*), or may not be observed. The algorithms above do not treat edge and non-edge information equally: Algorithms 1, 2 and 4 cannot report non-edges, and the deletion step in Algorithm 3 biases it towards edges over non-edges. However, the locations of non-edges can carry significant information.

Negative sampling schemes are "add-on" algorithms that are applied to the output of a graph sampling algorithm and augment it by non-edge information. Let $\overline{G}_k$ denote a sample generated by one of the algorithms above from an input graph $\overline{G}_n$.

Algorithm A (Negative sampling: Induced)**.**

(i) Report the subgraph induced by $\overline{G}_k$, in the input graph $\overline{G}_n$ from which $\overline{G}_k$ was drawn.

Another method, originating in language modeling [13, 6], is based on the unigram distribution: Define a probability distribution on the vertex set of $\overline{G}_k$ by $P_n(v) := \text{Prob}\{v \in \overline{H}_k\}$, the probability that v would occur in a separate, independent sample $\overline{H}_k$ generated from $\overline{G}_n$ by the same algorithm as $\overline{G}_k$. For $\tau > 0$, we define a distribution $P_n^\tau(v) := (P_n(v))^\tau / Z(\tau)$, where $Z(\tau)$ is the appropriate normalization.

Algorithm B (Negative sampling: Unigram)**.** For each vertex v in $\overline{G}_n$:

(i) Select k vertices $v_1, \ldots, v_k \overset{iid}{\sim} P_n^\tau$.

(ii) If (v, v_j) is a non-edge in $\overline{G}_n$, add it to $\overline{G}_n$.

The canonical choice in the embeddings literature is $\tau = \frac{3}{4}$ [13].

5 Theory

We now turn to formalizing and establishing theoretical properties of relational ERM. Particularly, (i) relational ERM satisfies basic theoretical desiderata, and (ii) Sample should be viewed as a model component. We first give the results, and then discuss their interpretation and significance.

When the data is unstructured (i.e., no link structure), theoretical analysis of ERM relies on the assumption that the data is generated i.i.d. The i.i.d. assumption is ill-defined for relational data. Any analysis requires some analogous assumption for how the data $\overline{G}_n$ is generated. Following recent work emphasizing the role of sampling theory in modeling graph data [15, 24, 1, 5], we model $\overline{G}_n$ as a random sample drawn from some large population network. Specifically, we consider a population graph $\mathcal{G}$ with $|\mathcal{G}|$ edges, and assume that the observed sample $\overline{G}_n$ of size n is generated by p-sampling from $\mathcal{G}$, with $p = n/\sqrt{|\mathcal{G}|}$. We assume the population graph is "very large," in the sense that $|\mathcal{G}| \to \infty$. The distribution of $\overline{G}_n$ in the "infinite population" case is well-defined [1].

The analogy with i.i.d. data generation is two-fold: Foundationally, the i.i.d. assumption is equivalent to assuming the data is collected by uniform sampling from some population [17], and p-sampling is a direct analogue [24, 1, 15]. Pragmatically, both assumptions strike a balance between being flexible enough to capture real-world data [2, 23] and simple enough to allow precise theoretical statements.

We establish results for several choices of $\mathsf{Sample}(G_n, k)$. Edges may be selected by either p-sampling with $p = k/\sqrt{n}$—note the size of $\mathsf{Sample}(G_n, k)$ is free of n—or by using a simple random walk of length k (Algorithm 1 or Algorithm 2). Negative examples may be chosen by Algorithm A or Algorithm B.

The main result guarantees that the limiting risk of the parameter we learn depends only on the population and the model, and not on idiosyncrasies of the training data.

Theorem 5.1. *Suppose that G_n is collected by p-sampling as described above, that $k \in \mathbb{N}$ is fixed, and that* Sample *is fixed to a sampling algorithm based on either p-sampling or random walk sampling as described above. Suppose further that the parameter setting $\bar{\theta} = (\bar{\gamma}, \bar{\lambda})$ satisfies mild technical conditions given in the appendix. there is some constant $c_{\bar{\theta}}(\mathsf{Sample}, k) \in \mathbb{R}_+$ such that*

$$\hat{R}_k(\bar{\theta}; \overline{G}_n) \to c_{\bar{\theta}}(\mathsf{Sample}, k) \tag{6}$$

both in probability and in L_1 as $n \to \infty$. Moreover, there is some constant $c_*(\mathsf{Sample}, k) \in \mathbb{R}_+$ such that

$$\min_\theta \hat{R}_k(\theta; \overline{G}_n) \to c_*(\mathsf{Sample}, k) \tag{7}$$

both in probability and in L_1, as $n \to \infty$.

The limits depend on the choice of Sample *(and k), and usually do not agree between different sampling schemes.*

The result is proved for Sample based on p-sampling in Appendix C and for random-walk based sampling in Appendix D.

Classical ERM guarantees usually apply even to the parameter itself, not just its risk. In the relational setting, the possibly complicated interplay of the learned embeddings makes such results more difficult. The next two results build on Theorem 5.1 to establish (partial) guarantees for the parameter itself.

We establish a convergence result for the global parameters output by a two-stage procedure where the embedding vectors are learned first. Such a result is applicable, for example, when predicting vertex attributes from embedding vectors that are pre-trained to explain graph structure. The proof is given in Appendix E.

Theorem 5.2. *Suppose the conditions of Theorem 5.1, and also that the loss function verifies a certain strict convexity property in γ, given explicitly in the appendix. Let $\tilde{\gamma}_n = \text{argmin}_\gamma \min_\lambda \hat{R}_k(\gamma, \lambda; \overline{G}_n)$. Then $\tilde{\gamma}_n \to \tilde{\gamma}_*(\mathsf{Sample}, k)$ in probability for some constant $\tilde{\gamma}_*(\mathsf{Sample}, k)$.*

We next establish a stability result showing that collecting additional data does not dramatically change

learned embeddings. The proof is given in Appendix F.

Theorem 5.3. *Suppose the conditions of Theorem 5.1, and also that the loss function is twice differentiable and the Hessian of the empirical risk is bounded. Let* $\hat{\lambda}_{n+1}|_n$ *denote the restriction of the embeddings* $\hat{\lambda}_{n+1}$ *to the vertices present in* G_n*. Then* $\hat{\lambda}_n - \hat{\lambda}_{n+1}|_n \to 0$ *in probability, as* $n \to \infty$*.*

Interpretation and Significance

The properties we establish are minimal desiderata that one might demand of any sensible learning procedure. Nevertheless, such results have not been previously established for relational learning methods. The obstruction is the need for a suitable analogue of the i.i.d. assumption. The demonstration that population sampling can fill this role is itself a main contribution of the paper. Indeed, the results we establish are weaker than the analogous guarantees for classical ERM, and main significance is perhaps the demonstration that such results can be established at all. This is important both as a foundational step towards a full theoretical analysis of relational learning, and because it strengthens the analogy with classical ERM.

A question that highlights the need for additional theory is: how should observed data $\overline{G}_n$ be split into train and test sets? Performance on a test set should be indicative of performance on freshly sampled data. For unstructured data, fresh data is idealized as a uniform sample from a population. This is simulated by uniformly sampling the observed data to create a test set. There is not an established analogue in the relational setting. Thus, there is no canonical procedure for comparing performance of relational data models. It is unclear in general how to empirically validate relational models without appeal to theory.

A strength of our arguments is that they are largely agnostic to the particular choice of model, mitigating the need for model-specific analysis and justification. For example, our results include random-walk based graph representation methods as a special case, providing some post-hoc support for the use of such methods.

The limits in Theorems 5.1 and 5.2 depend on the choice of Sample. Accordingly, the limiting risk and learned parameters depend on Sample in the same sense they depend on the choice of predictor class and the loss function; i.e., Sample is a model component. This underscores the need to consider the choice in model design, either through heuristics—e.g., random-walk sampling upweights the importance of high degree vertices relative to p-sampling—or by trying several choices experimentally.

6 Experiments

The practical advantages of using relational ERM to define new, task-specific, models are: (i) Mini-batch SGD can be used in a plug-and-play fashion to solve the optimization problem. This allows inference to scale to large data. And, (ii) by varying Sample we may improve model quality. We have used relational ERM to define novel models in Section 3. The models are determined by (4) and (5) up to the choice of Sample. We now study these example models empirically.[5] The main observations are: (i) SGD succeeds in quickly fitting the models in all cases. And, (ii) the choice of Sample has a dramatic effect in practice. Additionally, we observe that the best model for the semi-supervised node classification task uses p-sampling. p-sampling has not previously been used in the embedding literature, and is very different from the random-walk based schemes that are commonly used.

Node classification problems

We begin with the semi-supervised node classification task described in Section 3, using the model Eq. (4) with different choices of Sample. We study the blog catalog and protein-protein interaction data reported in [7], summarized by the table to the right. We pre-process the data to remove self-edges, and restrict each network to the largest connected component. Each vertex in the graph is labeled, and 50% of the labels are censored at training time. The task is to predict these labels at test time.

	Blogs	Protein
Vertices	10,312	3,890
Edges	333,983	76,584
Label Dim.	39	50

Table 1: Average Macro-F1 for Two-Stage Training.

Choice of Sample	Alg. #	Blogs	Protein
rw/skipgram+ns	1+B	0.18	0.16
rw/induced+ind	2+A	0.08	0.08
rw/induced+ns	2+B	0.18	0.16
p-samp+ind.	3+A	0.17	0.14
p-samp+ns	3+B	0.22	0.16
unif. edge+ns	4+B	0.21	0.15

Two-stage training. We first train the model (4) using no label information to learn the embeddings (that is, with $q = 0$). We then fit a logistic regression to predict vertex features from the trained embeddings.

[5] Code in supplementary material.

Table 2: Average Macro-F1 for Simultaneous Training. Columns are labeled by the sampling scheme used to draw test vertices.

	Blog catalog			Protein-Protein		
Sample	Unif.	p-samp	rw	Unif.	p-samp	rw
rw/skipgram+ns (Alg. 1+B)	0.20	0.26	0.27	0.25	0.32	0.34
p-samp+ns (Alg. 3+B)	0.30	0.34	0.35	0.30	0.37	0.39
Node2Vec (reported)	0.26	-	-	0.18	-	-

This two stage approach is a standard testing procedure in the graph embedding literature, e.g. [16, 7]. We use the same scoring procedure as Node2Vec [7] (average macro F1 scores), and, where applicable, the same hyperparameters.

Table 1 shows the effect of varying the sampling scheme used to train the embeddings. As expected, we observe that the choice of sampling scheme affects the embeddings produced via the learning procedure, and thus also the outcome of the experiment. We further observe that sampling non-edges by unigram negative sampling gives better predictive performance relative to selecting non-edges from the vertex induced subgraph.

Simultaneous training. Next, we fit the model of Section 3 with $q = 0.001$—training the embeddings and global variables simultaneously. Recall that simultaneous training is enabled by the use of relational ERM. We choose label predictor π_γ as logistic regression, and adapt the label prediction loss to measure the loss only on vertices in the positive sample.

There is not a unique procedure for creating a test set for relational data. We report test scores for test-sets drawn according to several different sampling schemes. Results are summarized by Table 2. We observe:

- Simultaneous training improves performance.
- p-sampling outperforms the standard rw/skipgram procedure.
- This persists irrespective of how the test set is selected (i.e., it is not an artifact of the data splitting procedure).

Note that the average computed with uniform vertex sampling is the standard scoring procedure used in the previous table.

Wikipedia Category Embeddings

We consider the task of discovering semantic relations between Wikipedia categories, as described in Section 3. In contrast with the previous example, this task is not standard and a wholly new model is required.

We define a relational ERM model by choosing category embedding dimension $k = 128$, the loss function L in (5), and Sample as 1+B, the skipgram random walk sampler with unigram negative sampling. The data $\overline{G}_n$ is the Wikipedia hyperlink network from [12], consisting of Wikipedia articles from 2011-09-01 restricted to articles in categories containing at least 100 articles. For each article, we observe the Wikipedia categories that article belongs to.

The challenge for this task is that the dataset is relatively large—about 1.8M nodes and 28M edges—and the model is unusual—embeddings are assigned to vertex attributes instead of the vertices themselves. We observe that SGD converges in about 90 minutes on a desktop computer equipped with a Nvidia Titan Xp GPU. Fig. 1 on page 5 visualizes example trained embeddings, which clearly succeed in capturing latent semantic structure.

7 Conclusion

Relational ERM is a generalization of ERM from i.i.d. data to relational data. The key ideas are introducing Sample as a component of model design, which defines an analogue of the empirical distribution, and using the assumption that the data is sampled from a population network as an analogue of the i.i.d. assumption. Relational ERM models are defined by a loss function, a predictor class, and a sampling scheme. These models can be fit automatically using SGD. Accordingly, relational ERM provides an easy method to specify and fit relational data models, as illustrated in Sections 3 and 6.

The results presented here suggest a number of directions for future inquiry. Foremost: what is the relational analogue of statistical learning theory? The theory derived in Section 5 establishes initial results. A more complete treatment may provide statistical guidelines for model development. Our results hinge critically on the assumption that the data is collected by p-sampling; it is natural to ask how other data-generating mechanisms can be accommodated. Similarly, it is natural to ask for guidelines for the choice of Sample.

References

[1] C. Borgs, J. T. Chayes, H. Cohn, and V. Veitch. *Sampling perspectives on sparse exchangeable graphs*. 2017. arXiv: 1708.03237.

[2] F. Caron and E. B. Fox. "Sparse graphs using exchangeable random measures". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79.5 (2017), pp. 1295–1366. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/rssb.12233.

[3] B. P. Chamberlain, J. Clough, and M. P. Deisenroth. "Neural Embeddings of Graphs in Hyperbolic Space". In: *ArXiv e-prints* (May 2017). arXiv: 1705.10359 [stat.ML].

[4] J. Chen, J. Zhu, and L. Song. "Stochastic Training of Graph Convolutional Networks with Variance Reduction". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by J. Dy and A. Krause. Vol. 80. Proceedings of Machine Learning Research. Stockholmsmässan, Stockholm Sweden: PMLR, 2018, pp. 942–950.

[5] H. Crane and W. Dempsey. *A framework for statistical network modeling*. 2015. arXiv: 1509.08185.

[6] Y. Goldberg and O. Levy. *word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method*. 2014. arXiv: 1402.3722.

[7] A. Grover and J. Leskovec. "Node2Vec: Scalable Feature Learning for Networks". In: *Proc. 22nd Int. Conference on Knowledge Discovery and Data Mining (KDD '16)*. ACM, 2016, pp. 855–864.

[8] W. L. Hamilton, R. Ying, and J. Leskovec. *Inductive Representation Learning on Large Graphs*. June 2017. arXiv: 1706.02216.

[9] W. L. Hamilton, R. Ying, and J. Leskovec. *Representation Learning on Graphs: Methods and Applications*. 2017. arXiv: 1709.05584.

[10] T. N. Kipf and M. Welling. "Variational Graph Auto-Encoders". In: *ArXiv e-prints* (Nov. 2016). arXiv: 1611.07308 [stat.ML].

[11] T. N. Kipf and M. Welling. "Semi-Supervised Classification with Graph Convolutional Networks". In: *ICLR*. 2017.

[12] C. Klymko, D. Gleich, and T. G. Kolda. *Using Triangles to Improve Community Detection in Directed Networks*. 2014. arXiv: 1404.5874.

[13] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. *Distributed Representations of Words and Phrases and their Compositionality*. 2013. arXiv: 1310.4546.

[14] T. Mikolov, K. Chen, G. Corrado, and J. Dean. *Efficient Estimation of Word Representations in Vector Space*. 2013. arXiv: 1301.3781.

[15] P. Orbanz. *Subsampling large graphs and invariance in networks*. 2017. arXiv: 1710.04217.

[16] B. Perozzi, R. Al-Rfou, and S. Skiena. "DeepWalk: Online Learning of Social Representations". In: *Proc. 20th Int. Conference on Knowledge Discovery and Data Mining (KDD '14)*. ACM, 2014, pp. 701–710.

[17] D. N. Politis, J. P. Romano, and M. Wolf. *Subsampling*. Springer, 1999.

[18] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. van den Berg, I. Titov, and M. Welling. "Modeling Relational Data with Graph Convolutional Networks". In: *The Semantic Web*. Ed. by A. Gangemi, R. Navigli, M.-E. Vidal, P. Hitzler, R. Troncy, L. Hollink, A. Tordai, and M. Alam. Cham: Springer International Publishing, 2018, pp. 593–607. ISBN: 978-3-319-93417-4.

[19] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[20] R. van den Berg, T. N. Kipf, and M. Welling. "Graph Convolutional Matrix Completion". In: *ArXiv e-prints* (June 2017). arXiv: 1706.02263 [stat.ML].

[21] V. Vapnik. "Principles of Risk Minimization for Learning Theory". In: *Advances in Neural Information Processing Systems 4*. 1992, pp. 831–838.

[22] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[23] V. Veitch and D. M. Roy. *The Class of Random Graphs Arising from Exchangeable Random Measures*. Dec. 2015. arXiv: 1512.03099.

[24] V. Veitch and D. M. Roy. "Sampling and Estimation for (Sparse) Exchangeable Graphs". In: (2016). arXiv: 1611.00843.

[25] Z. Yang, W. Cohen, and R. Salakhudinov. "Revisiting Semi-Supervised Learning with Graph Embeddings". In: *Proceedings of The 33rd International Conference on Machine Learning*. Ed. by M. F. Balcan and K. Q. Weinberger. Vol. 48. Proceedings of Machine Learning Research. New York, New York, USA: PMLR, 2016, pp. 40–48.

Proofs of Theoretical Results for Empirical Risk Minimization and Stochastic Gradient Descent for Relational Data

A Overview of Proofs

The appendix is devoted to proving the theoretical results of the paper. These results are obtained subject to the assumption that the data is collected by p-sampling. This assumption is natural in the sense that it provides a reasonable middle ground between a realistic data collection assumption—p-sampling can result in complex models capturing many important graph phenomena [3, 6, 1]—and mathematical tractability—we are able to establish precise guarantees.

The appendix is organized as follows. We begin by recalling the connection between p-sampling and *graphex processes* in Appendix B.1; this affords a useful explicit representation of the data generating process. In Appendix B.2, we recall the method of exchangeable pairs, a technical tool required for our convergence proofs. Next, in Appendix B.3, we collect the necessary notation and definitions. Empirical risk convergence results for p-sampling are then proved in Appendix C and results for the random-walk in Appendix D. Finally, convergence results for the global parameters are established in Appendix E.

B Preliminaries

B.1 Graphex processes

Recall the setup for the theoretical results: we consider a very large population network P_t with t edges, and we study the graph-valued stochastic process $(G_n^t)_{n\in[0,\sqrt{t})}$ given by taking each G_n^t to be an $n/\sqrt{t}$-sample from P_t and requiring these samples to cohere in the obvious way. We idealize the population size as infinite by taking the limit $t \to \infty$. The limiting stochastic process $(G_n)_{n\in\mathbb{R}_+}$ is well defined, and is called a *graphex process* [2].

Graphex processes have a convenient explicit representation in terms of (generalized) *graphons* [6, 1, 3].

Definition B.1. A *graphon* is an integrable function $W : \mathbb{R}_+^2 \to [0, 1]$.

Remark B.2. This notion of graphon is somewhat more restricted than graphons (or graphexes) considered in full generality, but it suffices for our purposes and avoids some technical details.

We now describe the generative model for a graphex process with graphon W. Informally, a graph is generated by (i) sampling a collection of vertices $\{\nu_i\}$ each with latent features x_i, and (ii) randomly connecting each pair of vertices with probability dependent on the latent features. Let

$$\Pi = \{\eta_i\}_{i\in\mathbb{N}} = \{(\nu(\eta_i), x(\eta_i))\}_{i\in\mathbb{N}}$$

be a Poisson (point) process on $\mathbb{R}_+ \times \mathbb{R}_+$ with intensity $\Lambda \otimes \Lambda$, where Λ is the Lebesgue measure. Each atom of the point process is a candidate vertex of the sampled graph; the $\{\nu_i\}$ are interpreted as (real-valued) labels of the vertices, and the $\{x_i\}$ as latent features that explain the graph structure. Each pair of points (η_i, η_j) with $i \le j$ is then connected independently according to

$$1[(\eta_i, \eta_j) \text{ connected}] \overset{ind}{\sim} \text{Bern}(W(x_i, x_j)).$$

This procedure generates an infinite graph. To produce a finite sample of size n, we restrict to the collection of edges $\Gamma_n = \{(\eta_i, \eta_j) \ : \ \eta_i, \eta_j \le n\}$. That is, we report the subgraph induced by restricting to vertices with label less than n, and removing all vertices that do not connect to any edges in the subgraph. This last step is critical; in general there are an infinite number of points of the Poisson process such that $\eta_i < n$, but only a finite number of them will connect to any edge in the induced subgraph.

Modeling G_n as collected by p-sampling is essentially equivalent to positing that G_n is the graph structure of Γ_n generated by some graphon W. Strictly speaking, the p-sampling model induces a slightly more general generative model that allows for both isolated edges that never interact with the main graph structure, and for infinite star structures; see [2]. Throughout the appendix, we ignore this complication and assume that the dataset graph is generated by some graphon. It is straightforward but notationally cumbersome to extend this assumption to p-sampling in full generality.

B.2 Technical Background: Exchangeable Pairs

We will need to bound the deviation of the (normalized) degree of a vertex from its expectation. To that end, we briefly recall the method of exchangeable pairs; see [4] for details.

Definition B.3. A pair of real random variables (X, X') is said to be exchangeable if

$$(X, X') \overset{d}{=} (X', X).$$

Let $f : \mathbb{R} \to \mathbb{R}$ and $F : \mathbb{R}^2 \to \mathbb{R}$ be measurable function such that:

$$\mathbb{E}(F(X, X')|X) \overset{a.s}{=} f(X), \text{ and } \mathrm{F(X, X')} = -\mathrm{F(X', X)}.$$

Let

$$v(X) \triangleq \frac{1}{2}\mathbb{E}\Big(\big(f(X) - f(X')\big)F(X, X')\Big|X\Big),$$

and suppose that $|v(X)| \overset{a.s}{\leq} C$ for some $C \in \mathbb{R}$. Then

$$\forall x > 0,\ P(|f(X) - \mathbb{E}(f(X))| \geq x) \leq 2e^{-\frac{x^2}{2C}}.$$

Further, for all $p > 1$ and $x > 0$ it holds that:

$$P(|f(X) - \mathbb{E}(f(X))| > x) \leq \frac{(2p-1)^p \|v(X)\|_p^p}{x^p}.$$

B.3 Notation

For convenient reference, we include a glossary of important notation.

First, notation to refer to important graph properties:

- $\Pi = \{\eta_i = (\nu(\eta_i), x(\eta_i))\}$ is the latent Poisson process that defines the graphex process in Appendix B.1. The labels are ν and the latent variables are x.
- $\Pi_n \triangleq \Pi \cap [0, n] \times \mathbb{R}^+$ is the restriction of the Poisson process to atoms with labels in $[0, n]$.
- To build the graph from the point of process Π_n we need to introduce a process of independent uniform variables. Let
$$\mathbb{U}_\Pi \triangleq (U_{\eta_i, \eta_j})_{\eta_i, \eta_j \in \Pi}$$
be such that $\mathbb{U}_\Pi | \Pi$ is an independent process where $U_{\eta_1, \eta_2} \mid \Pi \overset{iid}{\sim} \mathrm{Uni}(0, 1)$
- $\Gamma_n \subset \mathbb{R}_+^2$ is the (random) edge set of the graphex process at size n.
- $V(\Gamma_n) \subset \mathbb{R}_+$ is the set of vertices of Γ_n.
- $\bar{\Gamma}_n = \{(\eta_i, \eta_j) \ : \ \eta_i, \eta_j \in V(\Gamma_n) \text{ and } (\eta_i, \eta_j) \notin \Gamma_n\}$ is all pairs of points in Γ_n that are not connected by an edge.
- The number of edges in the graph is $E_n = |\Gamma_n|$
- The neighbors of η in Γ_n are
$$\mathcal{N}_n(\eta) \triangleq \{\eta' \ : \ (\eta, \eta') \in \mathcal{P}_1(\Gamma_n)\}$$

- For all k, the set of paths of length k in Γ_n is

$$\mathcal{P}_k(\Gamma_n) \triangleq \{(\eta_i)_{i\leq k+1} \in V(\Gamma_n)^{k+1} \ : \ (\eta_i, \eta_{i+1}) \in \Gamma_n \ \forall i \leq k\}.$$

- The degree of ν in Γ_n is $d_n(\eta)$.
- Asymptotically, the number of edges of a graphex process scales as n^2 [1]. Let $\mathcal{E} \in \mathbb{R}_+$ be the proportionality constant

$$\mathcal{E} \triangleq \lim_{n\to\infty} \frac{E_n}{n^2}.$$

Next, we introduce notation relating to model parameters. Treating the embedding parameters requires some care. The collection of vertices of the graph is a random quantity, and so the embedding parameters must also be modeled as random. For graphex processes, this means the embedding parameters depend on the latent Poisson process used in the generative model. To phrase a theoretical result, it is necessary to assume something about the structure of the dependence. The choice we make here is: the embedding parameters are taken to be markings of the Poisson process Π. In words, the embedding parameter of a vertex may depend on the (possibly latent) properties of that vertex, but the embeddings are independent of everything else.

- The collection of all possible parameters is:

$$\Omega_\theta^\Pi \triangleq \{(\lambda_\eta, \gamma)_{\eta\in\Pi} \ : \ \lambda_\eta \in \Omega_\theta \ \forall \eta \in \Pi \text{ and } \gamma \in \Omega_\gamma\}.$$

  Note that we attach a copy of the global parameter to each vertex for mathematical convenience.
- For all $\bar\theta \in \Omega_\theta^\Pi$, let $\lambda(\bar\theta)$ denote the projection on Ω_λ^Π and let $\gamma(\bar\theta)$ denote the projection on Ω_γ.
- The following concepts and notations are needed to build a marking of the Poisson process: Let $m(\cdot,\cdot)$ be a distributional kernel on $\mathbb{R}_+ \times \Omega_\theta$. We generate the marks according to a distribution $\mathcal{Q}_\theta^\Pi$ on Ω_θ^Π, conditional on Π, such that if $\bar\theta|\Pi \sim \mathcal{Q}_\theta^\Pi$ then:
  - $(\bar\theta_\eta)_{\eta\in\Pi}$ is an independent process
  - $\bar\theta_\eta|\Pi \sim m(x(\eta),\cdot)$ for all $\eta \in \Pi$
- Let $\bar\Pi_n(\theta) \triangleq (\Pi_n, \mathbb{U}|_{\Pi_n}, \theta|_n)$ the augmented object that carries information about both the graph structure $(\Pi_n, \mathbb{U}|_{\Pi_n})$ and the model parameters θ.

C Basic asymptotics for p-sampling

We begin by establishing the result for p-sampling, with $p = k/\sqrt{n}$ and the non-edges chosen by taking the induced subgraph. This is the simplest case, and is useful for the introduction of ideas and notation. We consider more general approaches to negative sampling in the next section, where it is treated in tandem with random walk sampling. The same arguments can be used to extend p-sampling to allow for, e.g., unigram negative sampling used in our experiments.

For all $\bar\theta \in \Omega_\theta^\Pi$, and all $\Gamma'_k \subset \Gamma$, let $L(\Gamma'_k, \bar\theta)$ denote the loss on Γ'_k where $\bar\theta$ is restricted to the embeddings (and global parameters) associated with Γ'_k.

Theorem C.1. *Let $\bar\theta$ a random variable taking value in Ω_θ^Π such that $\bar\theta \ |\Pi \sim \mathcal{Q}_\theta^\Pi$, for a certain kernel m, then there is some constant $c_m^{\text{ps}} \in \mathbb{R}_+$ such that if $\|\mathcal{L}\|_\infty < \infty$ then*

$$\hat R_k(\Gamma_n, \bar\theta) \to c_m^{\text{ps}}$$

both a.s. and in L_1, as $n \to \infty$.

Moreover there is some constant $c_*^{\text{ps}} \in \mathbb{R}_+$ such that

$$\min_\theta \hat R_k(\Gamma_n, \theta) \to c_*^{\text{ps}}$$

both a.s. and in L_1, as $n \to \infty$.

Proof. We will first prove the first statement. Let $\bar{\theta}|\Pi \sim \mathcal{Q}_\theta^\Pi$, let $\Gamma(\bar{\theta})$ be the edge set of $\bar{\Pi}(\bar{\theta})$, and let $\Gamma^n(\bar{\theta})$ be the partially labeled graph obtained from $\Gamma(\bar{\theta})$ by forgetting all labels in $[0,n)$ (but keeping larger labels and the embeddings θ). Let $\mathcal{F}_n(\bar{\theta})$ be the σ-field generated by $\Gamma^n(\bar{\theta})$. The critical observation is

$$\hat{R}_k(\Gamma_n, \bar{\theta}) = \mathbb{E}[L(\Gamma_k, \bar{\theta}) \mid \mathcal{F}_n(\bar{\theta})]. \tag{8}$$

The reason is that choosing a graph by k/n-sampling is equivalent uniformly relabeling the vertices in $[0,n)$ and restricting to labels less than k; averaging over this random relabeling operation is precisely the expectation on the righthand side.

By the reverse martingale convergence theorem we get that:

$$\hat{R}_k(\Gamma_n, \bar{\theta}) \xrightarrow{\text{a.s.}, L_1} \mathbb{E}[L(\Gamma_k, \bar{\theta}) \mid \mathcal{F}_\infty(\bar{\theta})],$$

but as $\mathcal{F}_\infty(\bar{\theta})$ is a trivial sigma-algebra we get the desired result.

We will now prove the second statement. Let Γ^n be the partially labeled graph obtained from Γ by forgetting all labels in $[0,n)$ and let $\mathcal{F}_n$ be the σ-field generated by Γ^n. Further, we denote the set of embeddings of the graph Γ^m by:

$$\Omega_\theta^{\Gamma^m} \triangleq \{(\lambda_{\mathcal{V},\gamma})_{\mathcal{V}\in\Gamma^m} : \forall \mathcal{V} \in V(\Gamma^m)\ \lambda_\mathcal{V} \in \Omega_\lambda, \gamma \in \Omega_\gamma\}.$$

We are now ready to state the proof. Let $m \leq n$, and observe that:

$$\mathbb{E}[\min_{\theta\in\Omega_\theta^{\Gamma^n}} \hat{R}_k(\Gamma_n, \theta) \mid \mathcal{F}_m] \leq \min_{\theta\in\Omega_\theta^{\Gamma^m}} \mathbb{E}[L(\Gamma_k, \theta) \mid \mathcal{F}_m] \tag{9}$$

$$= \min_{\theta\in\Omega_\theta^{\Gamma^m}} \hat{R}_k(\Gamma_n, \theta). \tag{10}$$

Thus, $(\min_{\theta\in\Omega_\theta^{\Gamma^n}} \hat{R}_k(\Gamma_n,\theta))_{n\in\mathbb{R}_+}$ is a supermartingale with respect to the filtration $(\mathcal{F}_n)_{n\in\mathbb{R}_+}$. Moreover, by assumption, the loss is bounded and thus so also is the empirical risk. Supermartingale convergence then establishes that $\min_{\theta\in\Omega_\theta^{\Gamma^n}} \hat{R}_k(\Gamma_n,\theta))$ converges almost surely and in L_1 to some random variable that is measureable with respect to $\mathcal{F}_\infty$. The proof is completed by the fact that $\mathcal{F}_\infty$ is trivial. □

D Basic asymptotics for random-walk sampling

In this section we establish the convergence of the relational empirical risk defined by the random walk. The argument proceeds as follows: We first recast the subsampling algorithm as a random probability measure, measurable with respect to the dataset graph Γ_n. Producing a graph according to the sampling algorithm is the same as drawing a graph according to the random measure. Establishing that the relational empirical risk converges then amounts to establishing that expectations with respect to this random measure converge; this is the content of Theorem D.8. To establish this result, we show in Lemma D.6 that sampling from the random-walk random measure is asymptotically equivalent to a simpler sampling procedure that depends only on the properties of the graphex process and not on the details of the dataset. We allow for very general negative sampling distributions in this result; we show that how to specialize to the important case of (a power of) the unigram distribution in Lemma D.7.

D.1 Random-walk Notation

We begin with a formal description of the subsampling procedure that defines the relational empirical risk. We will work with random subset of the Poisson process Π; these translate to random subgraphs of Γ in the obvious way. Namely, if the sampler selects $\eta_i = (\nu_i, x_i)$ in the Poisson process, then it selects η_i in Γ.

Sampling follows a two stage procedure: we choose a random walk, and then augment this random walk with additional vertices—this is the negative-sampling step. The following introduces much of the additional notation we require for this section.

Definition D.1 (Random-walk sampler)**.** Let μ_n be a (random) probability measure over $V(\Gamma_n)$. Let $H = (\eta_i)_{i\leq M} = (\nu(\eta_i), \lambda(\eta_i))_{i\leq M}$ be a sequence of vertices sampled according to:

1. (random-walk) $\eta_1 \sim \frac{d_n(\eta_1)}{2E_n}$ and let $\eta_i|\eta_{i-1} \sim \text{unif}(\mathcal{N}_{\text{n}}(\eta_{\text{i}-1}))$ for $i \in (2, \ldots, r+1)$.
2. (augmentation) $\eta_{r+2:M}$ be a sequence of additional vertices sampled from μ_n independently from each other and also from $(\eta_1, \ldots, \eta_{r+1})$.

Let G_H be the vertex induced subgraph of Γ_n. Let $P_n = \mathbb{P}(G_H \in \cdot \mid \bar{\Pi}_n(\bar{\theta}))$ be the random probability distribution over subgraphs induced by this sampling scheme.

With this notation in hand, We rewrite the loss function and the risk in a mathematically convenient form

Definition D.2 (Loss and risk)**.** The loss on a subsample is

$$L(G_H, \bar{\theta}) \in [0, 1],$$

where we implicity restrict to the embeddings (and global parameters) associated with vertices in G_H. The empirical risk is

$$\mathbb{E}_{P_n}[L(G_H, \bar{\theta})|\bar{\Pi}_n(\bar{\theta})].$$

Remark D.3. Note that the subgraphs produced by the sampling algorithm explicitly include all edges and non-edges of the graph. However, the loss may (and generally will) depend on only a subset of the pairs. In this fashion, we allow for the practically necessary division between negative and positive examples. Skipgram augmentation can be handled with the same strategy.

We impose a technical condition on the distribution that the additional vertices are drawn from. Intuitively, the condition is that the distribution is not too sensitive to details of the dataset in the large data limit.

Definition D.4 (Augmentation distribution)**.** We say μ_n is an *asymptotically exchangeable augmentation distribution* if is there is a μ such that

- There is a deterministic function f such that $\mu(\eta) = f(x(\eta))$
- $\|\mu_n(\cdot) - \frac{\mu(\cdot)\mathbb{I}(\cdot \in \Gamma_n)}{n\ Z_n}\|_{TV} \xrightarrow{p} 0$, where $Z_n \triangleq \frac{1}{n}\sum_{\eta \in \Pi_n} \mu(\eta)$.

Lemma D.7 establishes that the unigram distribution respects these conditions.

D.2 Technical lemmas

We begin with some technical inequalities controlling sums over the latent Poisson process. To interpret the theorem, note that the degree of a vertex with latent property y is given by $f_n(y, \Pi)$ in the theorem statement.

Lemma D.5. *Let $(U_{x(\eta)})_{\eta \in \Pi}$ be such that $(U_{x(\eta)})_{\eta \in \Pi}|\Pi$ is distributed as a process of independent uniforms in $[0, 1]$ and let*

$$f_n(y, \Pi) \triangleq \sum_{\eta \in \Pi_n} \mathbb{I}(U_{x(\eta)} \leq W(y, x)),$$

for all $y \in \mathbb{R}_+$. Then the following hold:

1. *$\forall y \in \mathbb{R}_+$ such that $W(y, \cdot) \geq n^{-1+\frac{\epsilon}{4}}$, there are $p, K > 0$ such that $\forall \beta > 0$,*

$$\mathbb{P}(\Big|\frac{f_n(y, \Pi)}{nW(y, \cdot)} - 1\Big| \geq \beta) \leq \frac{K}{n^3\beta^p}.$$

2. *$\forall p > 0,\ \exists K_p$ such that $\forall \beta > 0$*

$$\mathbb{P}(\Big|\frac{f_n(y, \Pi))}{n} - W(y, \cdot)\Big| \geq \beta) \leq \frac{K_p}{n^p\beta^{2p}}$$

and

$$\mathbb{P}(\Big|\frac{E_n}{n^2\mathcal{E}} - 1\Big| \geq \beta) \leq \frac{K_p}{n^p\beta^{2p}}.$$

3. $\exists K \in \mathbb{R}_+$ *such that* $\forall y \in \mathbb{R}_+$ *such that* $W(y,\cdot) \leq n^{-1+\frac{\epsilon}{4}}$ *then* $\mathbb{P}(f_n(\Pi, y) \geq n^{\frac{\epsilon}{2}}) \leq \frac{K}{n^3}$.

Proof. We will first write the proof of the first statement, which is harder. We then highlight the differences in the other cases. We use the Stein exchangeable pair method, recalled in Appendix B.2.

Let $F : \mathbb{R}^2 \to \mathbb{R}$ be such that

$$\forall x, y \; F(x,y) = [x - y].$$

Let $\bar{J} \sim \text{unif}(\{0, \text{n} - 1\})$ and let

$$\Pi' = T_{[\bar{J},\bar{J}+1],[n,n+1]} \cdot \Pi_\nu \times \Pi_x,$$

where $T_{[\bar{J},\bar{J}+1],[n,n+1]}$ is the permutation of $[\bar{J}, \bar{J}+1]$ and $[n, n+1]$ and

$$T_{[\bar{J},\bar{J}+1],[n,n+1]} \cdot \Pi_\nu \times \Pi_x \triangleq \{(T_{[\bar{J},\bar{J}+1],[n,n+1]}(\nu), x), \; \forall (\nu, x) \in \Pi\}$$

Then we can check the following:

- As $\Pi \cap [0,n] \setminus [\bar{j}, \bar{j}+1] \times \mathbb{R}^+ = \Pi' \cap [0,n] \setminus [\bar{j}, \bar{j}+1] \times \mathbb{R}^+$ we obtain that

$$\begin{aligned}
&\mathbb{E}(\frac{f_n(y,\Pi)}{W(y,\cdot)} - \frac{f_n(y,\Pi')}{W(y,\cdot)} \big| \Pi_n) \\
&\overset{(a)}{=} \frac{1}{nW(y,\cdot)} [\sum_{j=0}^{n-1} \sum_{\Pi_{j+1} \setminus \Pi_j} \mathbb{I}(U_{x(\eta)} \leq W(y,x)) - \mathbb{E}(\mathbb{I}(U_{x(\eta)} \leq W(y,x)))] \\
&\overset{(b)}{=} \frac{f_n(y,\Pi)}{nW(y,\cdot)} - 1
\end{aligned}$$

  where (a) is obtained by complete independence of Π and where to get (b) we use the fact that (see [6])

$$\sum_{(\nu,x) \in \Pi_{j+1} \setminus \Pi_j} \mathbb{I}(U_{x(\eta)} \leq W(y,x)) \sim \text{Poi}(W(y,\cdot))$$

- Moreover, we can very similarly see that:

$$\begin{aligned}
&\Big\| \frac{1}{2n} \mathbb{E}([\frac{f_n(y,\Pi)}{W(y,\cdot)} - \frac{f_n(y,\Pi')}{W(y,\cdot)}]^2 \big| \Pi_n) \Big\|_p \\
&\leq \frac{1}{n^2 W(y,\cdot)^2} \Big\| \sum_{j=0}^{n-1} [[\sum_{(\nu,x) \in \Pi_{j+1} \setminus \Pi_j} \mathbb{I}(U_{x(\eta)} \leq W(y,x))]^2 + 2W(y,\cdot)] \Big\|_p \\
&\leq \frac{1}{n^2 W(y,\cdot)^2} \sum_{j=0}^{n-1} \Big\| [\sum_{(\nu,x) \in \Pi_{j+1} \setminus \Pi_j} \mathbb{I}(U_{x(\eta)} \leq W(y,x))]^2 \Big\|_p + 2W(y,\cdot) \\
&\leq \frac{C}{nW(y,\cdot)},
\end{aligned}$$

  where C is a constant that does not depend on n or y.

Therefore using the exchangeable pair method presented earlier and setting $p \geq \frac{12}{\epsilon}$ for all y such that $W(y,\cdot) \geq n^{\frac{\epsilon}{4}-1}$ we get that there is K,p such that for all $\epsilon > 0$

$$P(|\frac{\sum_{(\nu,x) \in \Pi_n} \mathbb{I}(U_{x(\eta)} \leq W(y,x))}{W(y,\cdot)} - 1| \geq \beta) \leq \frac{K}{n^3 \beta^p},$$

QED.

For the second statement, instead of $\frac{f_n(y,\Pi)}{W(y,\cdot)}$ we are interested in $f_n(y,\Pi)$, which is easier to handle. Indeed, using the same exchangeable pair (Π, Π') we get that:

- As $\Pi \cap [0,n] \setminus [\bar{j}, \bar{j}+1] \times \mathbb{R}^+ = \Pi' \cap [0,n] \setminus [\bar{j}, \bar{j}+1] \times \mathbb{R}^+$ we obtain that

$$\begin{aligned}&\mathbb{E}(f_n(y,\Pi) - f_n(y,\Pi')\big|\Pi_n)\\ &= \frac{1}{n} f_n(y,\Pi) - W(y,\cdot).\end{aligned}$$

- Moreover we can very similarly see that:

$$\begin{aligned}&\Big\| \frac{1}{2n}\mathbb{E}([f_n(y,\Pi) - f_n(y,\Pi')]^2\big|\Pi_n)\Big\|_p\\ &\leq \frac{1}{n^2}\sum_{j=0}^{n-1} \Big\|[\sum_{(\nu,x)\in\Pi_{j+1}\setminus\Pi_j} \mathbb{I}(U_{x(\eta)} \leq W(y,x))]^2\Big\|_p + 2W(y,\cdot)\\ &\leq \frac{C}{n},\end{aligned}$$

where C is a constant that does not depend on n or y. Therefore we get the desired result QED.

A very similar roadmap can be followed for E_n.

The last statement is a simple consequence of the preceding results. Indeed, for all $y \in \mathbb{R}$,

$$P(W(y) \leq n^{-1+\frac{\epsilon}{4}}, f_n(\Pi, y) \geq n^{\frac{\epsilon}{2}}) \leq P(|\frac{f_n(\Pi,y)}{n} - W(y,\cdot)| \geq n^{-\frac{\epsilon}{4}}) \leq \frac{K_{\frac{3}{1+\frac{\epsilon}{4}}}}{n^3}.$$

□

With this in hand, we establish the asymptotic equivalence of random-walk sampling and a sampling scheme that does not depend on the details of the dataset. This is the main component of the proof. Recall the notation introduced in Appendix D.1.

Lemma D.6. *Suppose that there is $\epsilon \in (0,1)$ such that the graphon W verifies*

$$W(x,\cdot) = O(x^{-1-\epsilon}).$$

Suppose further that the augmented sampling distributions $(\mu_n)_n$ satisfy the conditions of Definition D.4. Then, writing

$$P_n(H) \triangleq \mathbb{I}(\eta_{1:r+1} \in \mathcal{P}_r(\Pi_n)) \frac{\prod_{l=r+2}^{M} \mu_n(\eta_l)}{2N_e^n \prod_{i=2}^{r} d_n(\eta_i)}$$

and

$$\tilde{P}_n(H) \triangleq \mathbb{I}(\eta_{1:r+1} \in \mathcal{P}_r(\Pi_n)) \frac{\prod_{l=r+2}^{M} \mu(\eta_l)}{2n^M \mathcal{E} \prod_{i=2}^{r} W(x(\eta_i),\cdot)},$$

it holds that

$$\sup_{\bar{\theta}\in\Omega_{\bar{\theta}}^{\Pi}} \Big|\mathbb{E}_{P_n}\big(L(G_H,\bar{\theta})|\bar{\Pi}_n(\bar{\theta})\big) - \mathbb{E}_{\tilde{P}_n}\big(L(G_H,\bar{\theta})|\bar{\Pi}_n(\bar{\theta})\big)\Big| = o_p(1).$$

Proof. We can first see by the triangle inequality that if we write the following two measures:

$$P_n^*(H) \triangleq \mathbb{I}(\eta_{1:r+1} \in \mathcal{P}_r(\Pi_n)) \frac{\prod_{l=r+2}^{M} \mu(\eta_l)}{2N_e^n n^{M-(r+1)} \prod_{i=2}^{r} d_n(\eta_i)}$$

and

$$\tilde{P}_n^*(H) \triangleq \mathbb{I}(\eta_{1:r+1} \in \mathcal{P}_r(\Pi_n)) \frac{\prod_{i=2}^{r} \mathbb{I}(W(x(\eta_i),\cdot) \geq n^{-1+\frac{\epsilon}{4}}) \prod_{l=r+2}^{M} \mu(\eta_l)}{2n^M \mathcal{E} \prod_{i=2}^{r} W(x(\eta_i),\cdot)}$$

Then $\forall \beta > 0$:

$$
\begin{aligned}
&P\Big(\sup_{\bar{\theta} \in \Omega_{\theta}^{\Pi}} \Big| \mathbb{E}_{P_n}\big(L(G_H, \bar{\theta}) | \bar{\Pi}_n(\bar{\theta})\big) - \mathbb{E}_{\tilde{P}_n}\big(L(G_H, \bar{\theta}) | \bar{\Pi}_n(\bar{\theta})\big) \Big| > \beta) \\
&\leq P\Big(\sup_{\bar{\theta} \in \Omega_{\theta}^{\Pi}} \Big| \mathbb{E}_{P_n}\big(L(G_H, \bar{\theta}) | \bar{\Pi}_n(\bar{\theta})\big) - \mathbb{E}_{P_n^*}\big(L(G_H, \bar{\theta}) | \bar{\Pi}_n(\bar{\theta})\big) \Big| > \frac{\beta}{3}) \\
&+ P\Big(\sup_{\bar{\theta} \in \Omega_{\theta}^{\Pi}} \Big| \mathbb{E}_{P_n^*}\big(L(G_H, \bar{\theta}) | \bar{\Pi}_n(\bar{\theta})\big) - \mathbb{E}_{\tilde{P}_n^*}\big(L(G_H, \bar{\theta}) | \bar{\Pi}_n(\bar{\theta})\big) \Big| > \frac{\beta}{3}) \\
&+ P\Big(\sup_{\bar{\theta} \in \Omega_{\theta}^{\Pi}} \Big| \mathbb{E}_{\tilde{P}_n^*}\big(L(G_H, \bar{\theta}) | \bar{\Pi}_n(\bar{\theta})\big) - \mathbb{E}_{\tilde{P}_n}\big(L(G_H, \bar{\theta}) | \bar{\Pi}_n(\bar{\theta})\big) \Big| > \frac{\beta}{3}),
\end{aligned}
$$

therefore proving that the last terms converge to zero for any $\beta > 0$ is sufficient.

First we will prove that

$$
\sup_{\bar{\theta} \in \Omega_{\theta}^{\Pi}} \Big| \mathbb{E}_{P_n}\big(L(G_H, \bar{\theta}) | \bar{\Pi}_n(\bar{\theta})\big) - \mathbb{E}_{P_n^*}\big(L(G_H, \bar{\theta}) | \bar{\Pi}_n(\bar{\theta})\big) \Big| = o_p(1).
$$

Indeed, noting that,

$$
P_{n,i}^*(H) \triangleq \mathbb{I}(\eta_{1:r+1} \in \mathcal{P}_r(\Pi_n)) \frac{\prod_{l=r+2}^{r+1+i} \mu(\eta_l) \prod_{r+2+i}^{M} \mu_n(\eta_l)}{2E_n n^i \prod_{i=2}^{r} d_n(\eta_i)},
$$

it holds $\forall \beta > 0$ that

$$
\begin{aligned}
&P\Big(\sup_{\bar{\theta} \in \Omega_{\theta}^{\Pi}} \Big| \mathbb{E}_{P_n}\big(L(G_H, \bar{\theta}) | \bar{\Pi}_n(\bar{\theta})\big) - \mathbb{E}_{P_n^*}\big(L(G_H, \bar{\theta}) | \bar{\Pi}_n(\bar{\theta})\big) \Big| > \beta\Big) \\
&\overset{(a)}{\leq} \sum_{i=1}^{M} P\Big(\sup_{\bar{\theta} \in \Omega_{\theta}^{\Pi}} \big| \mathbb{E}_{P_{n,i}^*}\big(L(G_H, \bar{\theta})\big) - \mathbb{E}_{P_{n,i-1}^*}\big(L(G_H, \bar{\theta})\big) \big| > \frac{\beta}{M}\Big) \\
&\leq M P\big(\|\mu_n - \frac{\mu}{nZ_\mu}\|_{TV} > \frac{\beta}{\|L\|_\infty}\big).
\end{aligned}
$$

where (a) using telescopic sum. Therefore we have proven that the first element of the sum goes to 0.

Now we will prove that

$$
\sup_{\bar{\theta} \in \Omega_{\theta}^{\Pi}} \Big| \mathbb{E}_{P_n^*}\big(L(G_H, \bar{\theta}) \mid \bar{\Pi}_n(\bar{\theta})\big) - \mathbb{E}_{\tilde{P}_n^*}\big(L(G_H, \bar{\theta}) | \bar{\Pi}_n(\bar{\theta})\big) \Big| = o_p(1).
$$

For this we will want to approximate $\frac{n}{d_n(V_{u_i})}$ by $\frac{1}{W(u_i,\cdot)}$. However for this we need a good bound on $P(|\frac{d_n(V_{u_i})}{nW(u_i,\cdot)} - 1| \geq \epsilon)$. But this is possible only if $W(u_i, \cdot)$ is not too small.

Note that for all vertices $\eta \in \Pi_n$ if a path H passes through η at the i-th coordinate, for $i \geq 2$, then it means that there is only $d_n(\nu(\eta))$ possibilities for the $i-1$th vertex of the path. Therefore if $d_n(\nu(\eta))$ is small the probability that our random-walk passes through v, and is not the origin vertex, is asymptotically negligible.

Indeed for all $\eta \in \Pi_n$ such that $d_n(\nu(\eta)) \leq n^{\frac{\epsilon}{2}}$ it holds that for $k \geq 2$,

$$
P(\eta_i = \eta \mid \bar{\Pi}_n(\bar{\theta})) \leq \sum_{\eta' \in \Pi_n \cap \mathcal{N}_n(\eta)} P(\eta_{i-1} = \eta', \eta_i = \eta \mid \bar{\Pi}_n(\bar{\theta})) \overset{(*)}{\leq} \frac{n^{\frac{\epsilon}{2}}}{2N_n^e},
$$

where to get (*) we used the stationary property of the random walk.

Therefore we have:

$$
P(\min_{k \geq 2} d_n(\eta_k) \leq n^{-\frac{\epsilon}{2}} \mid \bar{\Pi}_n(\bar{\theta})) \leq \frac{r n^{\frac{\epsilon}{2}} \big| \{\eta \in \Pi_n, \text{ s.t. } 0 < d_n(\eta) \leq n^{\frac{\epsilon}{2}}\} \big|}{2N_n^e} \xrightarrow{p} 0,
$$

But we have that $\forall(\eta_i)_{i\leq r+1}$ s.t. $\forall i, W(x(\eta_i),\cdot)\geq n^{-1+\frac{\epsilon}{4}}$,

$$\begin{aligned}
&\Big|\frac{1}{2E_n\prod_{i=2}^r d_n(\eta_i)}-\frac{1}{2n^{r+1}\mathcal{E}\prod_{i=2}^r W(x(\eta_i),\cdot)}\Big|\\
&\overset{(a)}{\leq}\sum_{i=2}^r\frac{1}{2E_n n^{i-1}\prod_{l=2}^{r-i}d_n(\eta_l)\prod_{l=r-i+2}^r W(x(\eta_l),\cdot)}\Big|\frac{1}{d_n(\eta_{r-i+1})}-\frac{1}{nW(x(\eta_{r-i+1}),\cdot)}\Big|\\
&\quad+\frac{1}{n^{r-1}\prod_{l=2}^r W(x(\eta_l),\cdot)}\Big|\frac{1}{2N_e^n}-\frac{1}{2n^2\mathcal{E}}\Big|\\
&\leq\sum_{i=2}^r\frac{1}{2E_n n^{i-1}\prod_{l=2}^{r-i+1}d_n(\eta_l)\prod_{l=r-i+2}^r W(x(\eta_l),\cdot)}\Big|1-\frac{d_n(\eta_{r-i+1})}{nW(x(\eta_{r-i+1}),\cdot)}\Big|+\frac{1}{2n^{r-1}N_e^n\prod_{l=2}^r W(x(\eta_l),\cdot)}\Big|1-\frac{N_e^n}{n^2\mathcal{E}}\Big|,
\end{aligned}$$

where (a) comes from a simple telescopic sum re-writing.

Therefore if

$$\max_i\Big|1-\frac{d_n(\nu_i)}{nW(y_i,\cdot)}\Big|,\Big|1-\frac{N_e^n}{n^2\mathcal{E}}\Big|\leq\beta$$

then

$$\begin{aligned}
&\Big|\frac{1}{2E_n\prod_{i=2}^r d_n(\eta_i)}-\frac{1}{2n^{r+1}\mathcal{E}\prod_{i=1}^r W(x(\eta_i),\cdot)}\Big|\\
&\leq\beta\big[\sum_{i=2}^r\frac{1}{2E_n n^{i-1}\prod_{l=2}^{r-i+1}d_n(\eta_l)\prod_{l=r-i+2}^r W(x(\eta_l),\cdot)}+\frac{1}{2n^{r-1}N_e^n\prod_{l=2}^r W(x(\eta_l),\cdot)}\big]
\end{aligned}$$

Now note that for all i, and $\lambda'\in\Omega$

$$\begin{aligned}
&\sum_{\eta_{1:r+1}\in\mathcal{P}_r(\Pi_n)}\frac{\prod_{i=2}^r\mathbb{I}(W(x(\eta_i),\cdot)\geq n^{-1+\frac{\epsilon}{4}})}{2E_n n^{i-1}\prod_{l=2}^{r-i+1}d_n(\eta_l)\prod_{l=r-i+2}^r W(x(\eta_l),\cdot)}\mathbb{E}\big(L(G_H,\bar\theta)|\eta_{r+2:M_n},\Pi_n\big)\\
&\overset{(a)}{\leq}\sum_{\eta_{1:r}\in\mathcal{P}_{r-1}(\Pi_n)}d_n(\eta_r)\frac{\prod_{i=2}^r\mathbb{I}(W(x(\eta_i),\cdot)\geq n^{-1+\frac{\epsilon}{4}})}{2E_n n^{i-1}\prod_{l=2}^{r-i+1}d_n(\eta_l)\prod_{l=r-i+2}^r W(x(\eta_l),\cdot)}\mathbb{E}\big(L(G_H,\bar\theta)|\eta_{r+2:M_n},\Pi_n\big)\\
&\leq\|L\|_\infty\max_{y\in N_v^n(\Pi)\ \text{s.t. }W(y,\cdot)\geq n^{-1+\frac{\epsilon}{4}}}\frac{d_n(y)}{nW(y,\cdot)}\sum_{\eta_{1:r}\in\mathcal{P}_{r-1}(\Pi_n)}\frac{\prod_{i=2}^r\mathbb{I}(W(x(\eta_i),\cdot)\geq n^{-1+\frac{\epsilon}{4}})}{2E_n n^{i-1}\prod_{l=2}^{r-i+1}d_n(\eta_l)\prod_{l=r-i+2}^{r-1}W(x(\eta_l),\cdot)}
\end{aligned}$$

where (a) is a simple consequence form the fact that:

$$\mathrm{card}\{\eta\in\eta(\Pi_\mathrm{n},\mathrm{r})\text{ s.t. }\eta|_{1:\mathrm{r}}=(\nu_\mathrm{i},\mathrm{y_i})_{1:\mathrm{r}}\}=\mathrm{d_n}(\nu_\mathrm{r})\mathrm{card}\{\eta\in\eta(\Pi_\mathrm{n},\mathrm{r}-1)\text{ s.t. }\eta|_{1:\mathrm{r}-1}=(\nu_\mathrm{i},\mathrm{y_i})_{1:\mathrm{r}-1}\}.$$

Therefore, by induction, we can get that for all i

$$\begin{aligned}
&\sum_{\eta_{1:r+1}\in\mathcal{P}_r(\Pi_n)}\prod_{i=2}^r\mathbb{I}(W(x(\eta_i),\cdot)\geq n^{-1+\frac{\epsilon}{4}})\frac{\mathbb{E}\big(L(G_H,\bar\theta)|\eta_{r+2:M},\Pi_n\big)}{E_n n^{i-1}\prod_{l=2}^{r-i+1}d_n(\eta_l)\prod_{l=r-i+2}^r W(x(\eta_l),\cdot)}\\
&\leq r\|L\|_\infty\max_{y\in N_v^n(y)\ \text{s.t. }W(y,\cdot)\geq n^{-1+\frac{\epsilon}{4}}}|\frac{d_n(y)}{nW(y,\cdot)}-1|+\|L\|_\infty.
\end{aligned}$$

Therefore if we note

$$A_n(\beta)\triangleq\{\max_{y\in N_v^n(y)\ \text{s.t. }W(y,\cdot)\geq n^{-1+\frac{\epsilon}{4}}}|\frac{d_n(y)}{nW(y,\cdot)}-1|\leq\beta,|\frac{N_e^n}{n^2\mathcal{E}}-1|\leq\beta\}$$

Then we can see the following:

- On $A_n(\beta)$ we will have that as $\eta_{1:r+1}\perp\!\!\!\perp\eta_{r+2:M}$ using the result that we previously got we have that:

$$\sup_{\bar\theta\in\Omega_\theta^\Pi}\Big|\mathbb{E}_{P_n^*}\big(L(G_H,\bar\theta)|\bar\Pi_n(\bar\theta)\big)-\mathbb{E}_{\tilde P_n^*}\big(L(G_H,\bar\theta)|\bar\Pi_n(\bar\theta)\big)\Big|\leq(r+1)^2\|L\|_\infty\beta$$

- And in addition we know that there is $K_1, K_2 < \infty$ s.t

$$
\begin{aligned}
P(A_n(\beta)^c) &\leq P(|\frac{N_e^n}{n^2\mathcal{E}} - 1| \geq \beta) + \mathbb{E}\big(\sum_{\eta_{1:r+1} \in \mathcal{P}_r(\Pi_n)} \mathbb{I}(|\frac{d_n(y)}{nW(y,\cdot)} - 1| \geq \beta)\big) \\
&\overset{(a)}{\leq} P(|\frac{N_e^n}{n^2\mathcal{E}} - 1| \geq \beta) + n\int_{\mathbb{R}^+} \mathbb{I}(W(x,\cdot) \geq n^{-1+\frac{\epsilon}{4}}) P\big(|\frac{f_n(x,\Pi)}{nW(x,\cdot)} - 1| \geq \beta\big)\big) dx \\
&\overset{(b)}{\leq} \frac{K_1}{n\beta} + \frac{K_2}{\beta^p n^2}\int_{\mathbb{R}^+} \mathbb{I}(W(x,\cdot) \geq n^{-1+\frac{\epsilon}{4}}) dx \\
&\leq \frac{K_1}{n\beta} + \frac{K_2}{\beta^p n^2} n^{1-\frac{3\epsilon}{2+2\epsilon}} \to 0,
\end{aligned}
$$

  where (a) comes from Slivnyak–Mecke theorem and (b) from Lemma D.5.

  Thus, we have successfully proven that

$$
\sup_{\bar{\theta} \in \Omega_\theta^\Pi} \Big| \mathbb{E}_{P_n^*}\big(L(G_H, \bar{\theta}) | \bar{\Pi}_n(\bar{\theta})\big) - \mathbb{E}_{\tilde{P}_n^*}\big(L(G_H, \bar{\theta}) | \bar{\Pi}_n(\bar{\theta})\big) \Big| = o_p(1)
$$

  QED

Now we are going to prove the last part, i.e.

$$
\sup_{\bar{\theta} \in \Omega_\theta^\Pi} \Big| \mathbb{E}_{\tilde{P}_n^*}\big(L(G_H, \bar{\theta}) | \bar{\Pi}_n(\bar{\theta})\big) - \mathbb{E}_{\tilde{P}_n}\big(L(G_H, \bar{\theta}) | \bar{\Pi}_n(\bar{\theta})\big) \Big| = o_p(1)
$$

For this we can note that that for all $i \geq 2$

$$
\begin{aligned}
&\| \frac{1}{n^{r+1}} \sup_{\lambda' \in \Omega_\theta^\Pi} \sum_{\eta_{1:r+1} \in \mathcal{P}_r(\Pi_n)} \frac{\mathbb{I}(W(x(\eta_i),\cdot) < n^{-1+\frac{\epsilon}{4}})}{2n^{r+1}\mathcal{E}\prod_{i=2}^r W(x(\eta_i),\cdot)} \mathbb{E}\big(L(G_H, \bar{\theta}) | \bar{\Pi}_n(\bar{\theta}), \eta_{r+2,M}\big) \|_{L_1} \\
&\overset{(a)}{\leq} \|L\|_\infty \int_{\mathbb{R}^{r+1}} \mathbb{I}(W(x(\eta_i),\cdot) < n^{-1+\frac{\epsilon}{4}}) \frac{\prod_{j=1}^r W(x_j, x_{j+1})}{\prod_{j=2}^r W(x_j,\cdot)} dx_{1:r+1} \\
&\overset{(b)}{\leq} \|L\|_\infty \int_{\mathbb{R}^{i}} \mathbb{I}(W(x(\eta_i),\cdot) < n^{-1+\frac{\epsilon}{4}}) \frac{\prod_{j=1}^{i-1} W(x_j, x_{j+1})}{\prod_{j=2}^{i-1} W(x_j,\cdot)} dx_{1:i} \\
&\overset{(c)}{\leq} \|L\|_\infty \int_{\mathbb{R}} W(x(\eta_i),\cdot) \mathbb{I}(W(x(\eta_i),\cdot) < n^{-1+\frac{\epsilon}{4}}) dx_i \xrightarrow{n\to\infty} 0,
\end{aligned}
$$

where to get (a) we used both the fact that L was bounded and the independence of the uniforms; to get (b) we integrated coordinates $r+1$ to $i+1$ and used the following definition:

$$
\forall x \int W(x', x) dx' = W(x, \cdot).
$$

We similarly got (c) where instead we integrated the coordinates from 1 to $i-1$.

Therefore we have successfully proven that

$$
\sup_{\bar{\theta} \in \Omega_\theta^\Pi} \Big| \mathbb{E}_{\tilde{P}_n^*}\big(L(G_H, \bar{\theta}) | \bar{\Pi}_n(\bar{\theta})\big) - \mathbb{E}_{\tilde{P}_n}\big(L(G_H, \bar{\theta}) | \bar{\Pi}_n(\bar{\theta})\big) \Big| = o_p(1)
$$

Hence we have proven the desired results □

We now turn to the question of which augmentation distributions will satisfy the conditions of the previous result. We show that the conditions hold for any distribution defined by a differentiable function of the unigram distribution; in particular, this covers the unigram distribution to the power of $3/4$ that is used to define unigram negative sampling.

Lemma D.7. *Let $\eta_{1:r+1}$ be sampled by a random walk on G_n, and let the random-walk unigram distribution be defined by*

$$
\mathrm{Ug}_{\Gamma_n}(\eta) = \mathbb{P}(\exists i \leq r+1, \text{ s.t. } \tilde{\eta_i} = \eta \mid \bar{\Pi}_n(\bar{\lambda})).
$$

Suppose that μ_n is defined by

$$\mu_n(\eta) \propto \mathrm{Ug}_{\Gamma_n}(\eta)^\alpha,$$

for a certain $\alpha > 0$. Then, defining μ by

$$\mu(\eta) \propto (r+1)^\alpha \frac{W(x,\cdot)^\alpha}{\mathcal{E}^\alpha},$$

it holds that

$$\|\mu_n - \frac{\mu(\cdot)\mathbb{I}(\cdot \in \Pi_n)}{nZ_n}\|_{TV} \xrightarrow{p} 0$$

Proof. We will for simplicity prove the result for $\alpha = 1$, the other cases can be obtained following a similarly, although the computations are more involved.

First, self-intersections of the walk are asymptotically negligible:

$$\begin{aligned}
&\sum_{\eta\in\Pi_n} \big|P(\exists i \le r+1,\ \text{s.t. } \tilde{\eta}_i = \eta | \bar{\Pi}_n(\bar{\lambda})) - \sum_{i=1}^{r+1} P(\tilde{\eta}_i = \eta | \bar{\Pi}_n(\bar{\lambda}))\big| \\
&\overset{(a)}{\le} \sum_{\eta\in\Pi_n}\sum_{i=1}^{r+1} P(\tilde{\eta}_i = \eta|\bar{\Pi}_n(\bar{\lambda}))P(\exists j \in [i+1, r+1],\ \eta_j = \eta | \eta_i = \eta, \bar{\Pi}_n(\bar{\lambda})) \xrightarrow{P,(b)} 0,
\end{aligned}$$

where (b) comes from the dominated convergence theorem and (a) comes from the fact that for all η

$$\begin{aligned}
&\Big|\mathbb{E}\big(\mathbb{I}\big(\exists i \le r+1,\ \text{s.t. } \tilde{\eta}_i = \eta\big) - \sum_{i=1}^{r+1} \mathbb{I}(\tilde{\eta}_i = \eta)|\bar{\Pi}_n(\bar{\lambda})\big)\Big| \\
&\le \sum_{i=1}^{r+1} \mathbb{E}\big(\mathbb{I}(\tilde{\eta}_i = \eta,\ \exists j \ge i \text{ s.t. } \tilde{\eta}_j = \eta)|\Gamma_n\big)
\end{aligned}$$

Next, the limiting probability that a walk includes η is determined by its limiting relative degree $\frac{W(x(\eta),\cdot)}{2\mathcal{E}}$. To that end, we write:

$$\begin{aligned}
&\sum_{\eta\in\Pi_n} \big|\sum_{i=1}^{r+1} P(\tilde{\eta}_i = \eta|\bar{\Pi}_n(\bar{\lambda})) - \frac{(r+1)W(x(\eta),\cdot)}{2n\mathcal{E}}\big| \\
&\overset{(a)}{\le} \sum_{\eta\in\Pi_n} \big|\frac{(r+1)d_n(\eta)}{2E_n} - \frac{(r+1)W(x(\eta),\cdot)}{2n\mathcal{E}}\big|
\end{aligned}$$

where (a) comes from the stationarity proprieties of the simple random walk.Then, using Lemma D.5, we see that:

$$\sum_{\eta\in\Pi_n} \big|\sum_{i=1}^{r+1} P(\tilde{\eta}_i = \eta|\bar{\Pi}_n(\bar{\lambda})) - \frac{(r+1)W(x(\eta),\cdot)}{2n\mathcal{E}}\big| = o_p(1).$$

Finally,

$$\begin{aligned}
&\sum_{\eta\in\Pi_n} \big|\frac{(r+1)W(x(\eta),\cdot)}{2n\mathcal{E}}[1 - \frac{1}{\sum_{\eta\in\Pi_n}\frac{(r+1)W(x(\eta),\cdot)}{2n\mathcal{E}}}]\big| \\
&= \sum_{\eta\in\Pi_n} \frac{(r+1)W(x(\eta),\cdot)}{2n\mathcal{E}} - 1 \\
&= \sum_{\eta\in\Pi_n} \frac{(r+1)W(x(\eta),\cdot)}{2n\mathcal{E}} - P(\exists i \le r+1,\ \text{s.t. } \tilde{\eta}_i = \eta|\Gamma_n) = o_p(1).
\end{aligned}$$

□

D.3 Convergence for random walk sampling

Let $\bar{\theta}$ be a random element of Ω_θ^Π such that $\bar{\theta}|\Pi \sim \mathcal{Q}_\theta^\Pi$ for a certain kernel m. For brevity, we write

$$\hat{R}_k(G_n, \bar{\theta}) \triangleq \mathbb{E}_{P_n}\big(L(G_H, \bar{\theta})|\bar{\Pi}_n(\bar{\theta})\big).$$

for all $n \in \mathbb{R}_+$.

Theorem D.8. *There are constants $c_m^{\mathrm{rw}}, c_*^{\mathrm{rw}} \in \mathbb{R}_+$ such that*

$$\hat{R}_k(G_n, \bar{\theta}) \xrightarrow{p} c_m^{\mathrm{rw}},$$

and

$$\min_{\bar{\theta}\in\Omega_w^\Pi} \hat{R}_k(G_n, \bar{\theta}) \xrightarrow{p} c_*^{\mathrm{rw}}.$$

And those constants are respectively $\lim_n \mathbb{E}(\hat{R}_k(G_n, \bar{\theta}))$ *and* $\lim_n \mathbb{E}(\min_{\bar{\theta}\in\Omega_w^\Pi} \hat{R}_k(G_n, \bar{\theta}))$

Proof. Lemma D.6 states that

- $\mathbb{E}_{P_n}\big(L(G_H, \bar{\theta})|\bar{\Pi}_n(\bar{\theta})\big) - \mathbb{E}_{\tilde{P}_n}\big(L(G_H, \bar{\theta})|\bar{\Pi}_n(\bar{\theta})\big) = o_p(1)$.
- $\min_{\bar{\theta}\in\Omega_\theta^\Pi} \mathbb{E}_{P_n}\big(L(G_H, G_H(\lambda))|\bar{\Pi}_n(\bar{\theta})\big) - \min_{\bar{\theta}\in\Omega_\theta^\Pi} \mathbb{E}_{\tilde{P}_n}\big(L(G_H, G_H(\lambda))|\bar{\Pi}_n(\bar{\theta})\big) = o_p(1)$.

We will see that $\mathbb{E}_{\tilde{P}_n}$ inherits much of the nice distributional structure of the point process Π. This will be essential to the proof.

To see this we first define for all integers $i \in \mathbb{N}$ the restriction of the point process to points $\eta \in \Pi$ such that $\nu(\eta) \in (i, i+1]$,

$$\Pi|_{(i,i+1]} := \Pi_{i+1} \setminus \Pi_i.$$

And for all M sequence of integers $I = (I_1, \ldots, I_M) \in \mathbb{N}^M$ we write the following sequence of M restrictions of Π,

$$\Pi|_I \triangleq (\Pi|_{(I_1, I_1+1]}, \ldots, \Pi|_{(I_M, I_M+1]}).$$

This allows us to define the following M-dimensional array $X(\bar{\theta}) \triangleq (X_I(\bar{\theta}))_{I\in\mathbb{N}^M}$ where for all M integers $I = (I_1, \ldots, I_M) \in \mathbb{N}^M$,

$$X_I(\bar{\theta}) \triangleq \sum_{\eta_{1:M}\in\Pi|_I} \frac{\mathbb{I}(\eta_{1:r+1} \in \mathcal{P}_r(\Pi_n)) \prod_{l=r+2}^M \mu(x(\eta_l))}{2\mathcal{E}\prod_{i=2}^r W(x(\eta_i), \cdot)} L(G_H, G_H(\bar{\theta})).$$

This quantity is key as we can write that

$$\mathbb{E}_{\tilde{P}_n}\big(L(G_H, \bar{\theta})|\bar{\Pi}_n(\bar{\theta})\big) = \frac{1}{n^M} \sum_{i_{1:M}\leq n-1} X_{i_{1:M}}^{\bar{\theta}}. \tag{11}$$

Then using classical results on convergence of exchangeable arrays [5] we obtain that:

$$\mathbb{E}_{P_n}\big(L(G_H, \theta)|\bar{\Pi}_n(\bar{\theta})\big) \xrightarrow{p} \int_{\mathbb{R}_+^M} \mathcal{R}(x_{1:M}) \frac{\prod_{i=r+2}^M \mu(x_i)}{2\mathcal{E}\prod_{i=2}^r W(x_i, \cdot)} dx_{1:M},$$

where

$$\mathcal{R}(x_{1:M}) = \mathbb{E}\Big(L(G_{x_{1:M}}, G_{x_{1:M}}(\theta_{x_{1:M}})) \prod_{i=1}^r \mathbb{I}(U_i \leq W(x_i, x_{i+1}))\Big),$$

and where $G_{x_{1:M}}$ is the subgraph with vertices having intensities respectively $x_1, \ldots, x_m$, and $\forall i,\ \theta_{x_i} \overset{iid}{\sim} m(x_i, \cdot)$.

Now let write for all n, $\mathbb{F}_n$ the sigma-field of events invariant under joint permutations of the indexes in $[1, n]^M$. Then we can see that $(\min_{\bar{\theta}\in\Omega_\theta^{\Pi_n}} \frac{1}{\prod_{i=0}^{M-1}(n-i)} \sum_{I\in[|1,n-1|]^M} X_I(\bar{\theta}), \mathbb{F}_n)$ is a reverse supermartingale. Indeed

- $\min_{\bar{\theta}\in\Omega_\theta^{\Pi_n}} \frac{1}{\prod_{i=0}^{M-1}(n-i)} \sum_{I\in[|1,n-1|]^M} X_I(\bar{\theta})$ is $\mathbb{F}_n$ measurable as it is invariant under joint permutations of the indexes in $[1,n]^M$.
- For all $m \geq n$ let $\hat{\theta}_m \in \Omega_\theta^{\Pi}$ such that:

$$\sum_{I\in[|1,m-1|]^M} X_I(\hat{\theta}_m) = \min_{\bar{\theta}\in\Omega_\theta^{\Pi_m}} \sum_{I\in[|1,m-1|]^M} X_I(\bar{\theta})$$

Then we get

$$\begin{aligned}
&\mathbb{E}\Big(\min_{\bar{\theta}\in\Omega_\theta^{\Pi_n}} \frac{1}{n^M} \sum_{I\in[|1,n-1|]^M} X_I(\bar{\theta})|F_m\Big) \\
&\overset{(a)}{\leq} \mathbb{E}\Big(\frac{1}{n^M} \sum_{I\in[|1,n-1|]^M} X_I(\hat{\theta}_m)|F_m\Big) \\
&\overset{(b)}{\leq} \min_{\bar{\theta}\in\Omega_\theta^{\Pi_m}} \frac{1}{m^M} \sum_{I\in[|1,m-1|]^M} X_I(\bar{\theta}),
\end{aligned}$$

where (a) comes from Jensen and (b) comes from a standard argument in exchangeable arrays.

Therefore we have that:

$$\min_{\bar{\theta}\in\Omega_\theta^{\Pi}} \frac{1}{n^M} \sum_{I\in[|1,n-1|]^M} X_I(\bar{\theta}) - \mathbb{E}(\min_{\bar{\theta}\in\Omega_\theta^{\Pi}} \frac{1}{n^M} \sum_{I\in[|1,n-1|]^M} X_I(\bar{\theta})) \xrightarrow{p} 0.$$

□

E Convergence of global parameters

We now establish the second main convergence result. This result applies to the two stage procedure where the embeddings are learned first and the global parameters are then learned with the embeddings fixed. The result is that the learned global parameters will converge in the ordinary statistical consistency sense.

Our proof of this guarantee requires some technical conditions.

Definition E.1. Suppose that Ω_γ is a compact convex set. A loss function L is *ϵ-strictly convex* in γ if for all $\gamma, \gamma' \in \Omega_\gamma$, for all $\eta \in [0,1]$, and for all $\bar{\theta}_\gamma, \bar{\theta}_{\gamma'}, \bar{\theta}_{\eta\gamma'+(1-\eta)\gamma}$ such that

1. $\lambda(\bar{\theta}_\gamma) = \lambda(\bar{\theta}_{\gamma'}) = \lambda(\bar{\theta}_{\eta\gamma'+(1-\eta)\gamma})$, and
2. $\gamma(\bar{\theta}_\gamma) = \gamma,\ \gamma(\bar{\theta}_{\gamma'}) = \gamma',\ \gamma(\bar{\theta}_{(1-\eta)\gamma+\eta\gamma'}) = (1-\eta)\gamma + \eta\gamma'$

it holds that

$$L(G_H, \bar{\theta}_{\eta\gamma'+(1-\eta)\gamma}) \overset{\text{a.s.}}{<} \eta L(G_H, \bar{\theta}_{\gamma'}) + (1-\eta)L(G_H, \bar{\theta}_\gamma) - \epsilon.$$

Definition E.2. A loss function L is *uniformly continuous* if

$$\lim_{\gamma'\to\gamma} \Big\| \sup_{\bar{\lambda}\in\Omega_\lambda^{\Pi}} \big|L(G_H, \bar{\theta}_{\gamma'}) - L(G_H, \bar{\theta}_\gamma)\big| \Big\|_{L_1} = 0.$$

We write the risk as $\hat{R}_k(\gamma, \lambda; G_n)$.

Lemma E.3. *Suppose that there is $\epsilon > 0$ such that L is ϵ-strictly convex and uniformly continuous in γ, and that Ω_γ is a compact convex set. Let $(\hat{\gamma}_n)_n \in \Omega_\gamma^{\mathbb{N}}$ be a sequence of elements in Ω_γ such that, for all n,*

$$\min_{\lambda\in\Omega_\lambda^{\Pi}} \hat{R}_k(\hat{\gamma}_n, \lambda; G_n) = \min_{\gamma\in\Omega_\gamma} \min_{\lambda\in\Omega_\lambda^{\Pi}} \hat{R}_k(\gamma, \lambda; G_n).$$

Then

$$\hat{\gamma}_n \xrightarrow{p} \gamma^*,$$

where $\gamma^* = \text{argmin}_\gamma \text{lim}_\text{n}\, \mathbb{E}(\min_{\lambda\in\Omega_\lambda^{\Pi}} \hat{\text{R}}_\text{k}(\gamma, \lambda; \text{G}_\text{n}))$

Remark E.4. This result is valid for both random-walk and p-sampling.

Proof. Let $\hat{R}_k(\gamma; G_n) \triangleq \min_{\lambda \in \Omega_\theta^\Pi} \hat{R}_k(\gamma, \lambda; G_n)$.

Theorem D.8 for the random walk sampler and Theorem C.1 for p-sampling give the following for all γ:

$$\hat{R}_k(\gamma; G_n) - \mathbb{E}(\hat{R}_k(\gamma; G_n))) \xrightarrow{p} 0.$$

Let $(\hat{\gamma}_n)_n \in \Omega_\gamma^{\mathbb{N}}$ be a sequence such that

$$\hat{R}_k(\hat{\gamma}_n; G_n) = \min_{\gamma \in \Omega_\gamma} \hat{R}_k(\gamma; G_n).$$

Since $(\hat{\gamma}_n)_n$ is a sequence in the compact set Ω_γ there is a function $\phi : \mathbb{N} \to \mathbb{N}$ and $\tilde{\gamma}$ such that $\hat{\gamma}_{\phi(n)} \xrightarrow{d} \tilde{\gamma}$. But as Ω_γ is compact, an easy consequence of the covering lemma gives that:

$$\sup_{\gamma \in \Omega_\gamma} \left| \hat{R}_k(\gamma; G_n) - f(\gamma) \right| \xrightarrow{p} 0,$$

where $f : \gamma \to \lim_n \mathbb{E}(\hat{R}_k(\gamma; G_n))$. Therefore we have that

$$|\hat{R}_k(\hat{\gamma}_{\phi(n)}, G_{\phi(n)}) - f(\hat{\gamma}_{\phi(n)})| \xrightarrow{p} 0.$$

But using the expressions Eq. (11) and Eq. (8) derived in the proof of respectively Theorem D.8 and Theorem C.1 and the ϵ-strictly convex assumption on L we have that f is continuous and is strictly convex, and hence has a unique minimizer.

Therefore $\tilde{\gamma}$ must be deterministic equal to $\gamma^* \triangleq \operatorname{argmin}_\gamma f(\gamma)$. Indeed suppose by contradiction that it is not the case then there is $\eta > 0$ such that

$$\mathbb{P}(\hat{R}_k(\hat{\gamma}_{\phi(s)}, G_{\phi(s)}) - \hat{R}_k(\gamma^*, G_{\phi(s)}) > \eta) > \eta,$$

which is a contradiction of the definition of $(\hat{\gamma}_n)_n$. Therefore we have successfully proven that $\tilde{\gamma} = \gamma^*$.

And we have proved that $\hat{\gamma}_n \xrightarrow{p} \gamma^*$.

□

F Stability of embeddings

Theorem F.1. *Suppose the conditions of Theorem 5.1 (i.e., the form of* Sample*, that $\overline{G}_n$ is generated by a graphon process, and that parameter settings are markings of the latent Poisson process). Suppose that the loss function is twice differentiable and the Hessian of the empirical risk is bounded. Let $\hat{\lambda}_{n+1}|_n$ denote the restriction of the embeddings $\hat{\lambda}_{n+1}$ to the vertices present in G_n. Then $\hat{\lambda}_n - \hat{\lambda}_{n+1}|_n \to 0$ in probability, as $n \to \infty$.*

Proof. For notational simplicity, we consider the case with no global parameters and note that the same proof works if global parameters are included.

First, by a Taylor expansion about $\hat{\lambda}_n$,

$$\hat{R}_k(\hat{\lambda}_{n+1}|_n; \overline{G}_n) = \hat{R}_k(\hat{\lambda}_n; \overline{G}_n) + 0 + {}^1\!/_2(\hat{\lambda}_n - \hat{\lambda}_{n+1}|_n)^T H_n(\hat{\lambda}_n - \hat{\lambda}_{n+1}|_n),$$

where H_n is the Hessian evaluated at an appropriate point. Then, to prove the result it suffices to show that $\hat{R}_k(\hat{\lambda}_{n+1}|_n; \overline{G}_n) - \hat{R}_k(\hat{\lambda}_n; \overline{G}_n) \xrightarrow{p} 0$ as $n \to \infty$.

To that end, we first show $\hat{R}_k(\hat{\lambda}_{n+1}|_n; \overline{G}_n) \approx \hat{R}_k(\hat{\lambda}_{n+1}; \overline{G}_{n+1})$. By [1, Prop. 30], $E_n/n^2 \to \mathcal{E}$ a.s. as $n \to \infty$. Then, the expected number of edges selected by $\mathsf{Sample}(\overline{G}_{n+1}, k)$ that do not belong to $\overline{G}_n$ is:

$$k(1 - \mathbb{E}[e(\overline{G}_n)/e(\overline{G}_{n+1}) \mid \overline{G}_{n+1}] = o(1) \text{ a.s.} \tag{12}$$

We expand $\hat{R}_k(\hat{\lambda}_{n+1}; \overline{G}_{n+1})$ as:

$$\begin{aligned}\mathbb{E}[L(\mathsf{Sample}(\overline{G}_{n+1}, k); \hat{\lambda}_{n+1}) \mid \overline{G}_{n+1}] = \mathbb{E}[L(\mathsf{Sample}(\overline{G}_n, k); \hat{\lambda}_{n+1}|_n) \mid \overline{G}_n]\mathbb{P}(\mathsf{Sample}(\overline{G}_{n+1}, k) \subset \overline{G}_n \mid \overline{G}_{n+1}) \\ + \mathbb{E}[L(\mathsf{Sample}(\overline{G}_{n+1}, k); \hat{\lambda}_{n+1}) \mid \overline{G}_{n+1}]\mathbb{P}(\mathsf{Sample}(\overline{G}_{n+1}, k) \subsetneq \overline{G}_n \mid \overline{G}_{n+1}).\end{aligned} \tag{13}$$

By Markov's inequality and Eq. (12),

$$\mathbb{P}(\mathsf{Sample}(\overline{G}_{n+1}, k) \subsetneq \overline{G}_n \mid \overline{G}_{n+1}) \xrightarrow{p} 0,$$

as $n \to \infty$. By Theorem 5.1, $\mathbb{E}[L(\mathsf{Sample}(\overline{G}_{n+1}, k); \hat{\lambda}_{n+1}) \mid \overline{G}_{n+1}]$ converges to a constant in probability, so the second term of Eq. (13) converges to 0 in probability. Hence,

$$\hat{R}_k(\hat{\lambda}_{n+1}|_n; \overline{G}_n) - \hat{R}_k(\hat{\lambda}_{n+1}; \overline{G}_{n+1}) \xrightarrow{p} 0, \tag{14}$$

as $n \to \infty$.

By Theorem 5.1,

$$\hat{R}_k(\hat{\lambda}_n; \overline{G}_n) - \hat{R}_k(\hat{\lambda}_{n+1}; \overline{G}_{n+1}) \xrightarrow{p} 0, \tag{15}$$

as $n \to \infty$. The proof is completed by combining Eqs. (14) and (15). □

References

[1] C. Borgs, J. T. Chayes, H. Cohn, and N. Holden. *Sparse exchangeable graphs and their limits via graphon processes*. Jan. 2016. arXiv: 1601.07134.

[2] C. Borgs, J. T. Chayes, H. Cohn, and V. Veitch. *Sampling perspectives on sparse exchangeable graphs*. 2017. arXiv: 1708.03237.

[3] F. Caron and E. B. Fox. "Sparse graphs using exchangeable random measures". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79.5 (2017), pp. 1295–1366. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/rssb.12233.

[4] S. Chatterjee. "Concentration inequalities with exchangeable pairs (Ph.D. thesis)". In: *ArXiv Mathematics e-prints* (July 2005). eprint: math/0507526.

[5] O. Kallenberg. "Multivariate sampling and the estimation problem for exchangeable arrays". In: *Journal of Theoritical probability* (1999).

[6] V. Veitch and D. M. Roy. *The Class of Random Graphs Arising from Exchangeable Random Measures*. Dec. 2015. arXiv: 1512.03099.