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 supplement B.1; this affords a useful explicit representation of the data generating process. In supplement B.2, we recall the method of exchangeable pairs, a technical tool required for our convergence proofs. Next, in supplement B.3, we collect the necessary notation and definitions. Empirical risk convergence results for p-sampling are then proved in supplement C and results for the random-walk in supplement D. Convergence results for the global parameters are established in supplement E. Finally, in supplement F, we show that learned embeddings are stable in sense that they are not changed much by collecting a small amount of additional data.

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 our results 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

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

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

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

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 $|v(X)| \stackrel{a.s}{\leq} C$ 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 supplement 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 $U_{\eta_1,\eta_2} \mid \Pi \stackrel{iid}{\sim} \mathrm{Uni}(0, 1)$
- $\Gamma_n \subset \mathbb{R}_+^2$ 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 $c_*^{\mathrm{ps}} \in \mathbb{R}_+$ *such that*

$$\min_\theta \hat{R}_k(\Gamma_n, \theta) \to c_*^{\mathrm{ps}}$$

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 $m \le n$, and observe that:

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

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

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 $\eta_i|\eta_{i-1} \sim \mathrm{unif}(\mathcal{N}_{\mathrm{n}}(\eta_{\mathrm{i}-1}))$ 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 supplement 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:

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

  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 supplement 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

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

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

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

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:

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

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:

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

- 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

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

  QED

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

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

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

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

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:

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

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

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

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:

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

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,

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

□

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-strongly 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

$$L(G_H, \bar{\theta}_{\eta\gamma'+(1-\eta)\gamma}) \overset{\text{a.s.}}{<} \eta L(G_H, \bar{\theta}_{\gamma'}) + (1-\eta)L(G_H, \bar{\theta}_\gamma) - \frac{1}{2}\epsilon\eta(1-\eta)\|\gamma - \gamma'\|_2^2.$$

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

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 any 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-strongly 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 $\gamma^* \triangleq \text{argmin}_\gamma f(\gamma)$. 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.