A Introduction to exchangeable graph or graphon models

In the following discussion, we assume all graphs mentioned are undirected and have no self loops. A *graphon model* refers to a probabilistic model on a graph on a countable set $\mathcal{V} \subseteq \mathbb{N}$, defined via a *graphon*, which we define as a symmetric measurable function $W : [0,1]^2 \to [0,1]$. To define the law of the graph, for each vertex $u \in \mathcal{V}$, we assign an independent latent variable $\lambda_u \sim \mathrm{Unif}[0,1]$, and then assign edges independently according to the law

$$a_{uv} | \lambda_u, \lambda_v \sim \mathrm{Bernoulli}\big(W(\lambda_u, \lambda_v)\big) \tag{16}$$

independently for $u < v$, and then setting $a_{vu} = a_{uv}$ for $u > v$.

The Aldous-Hoover theorem [1] then gives the following equivalence between probabilistic models of a graph on a vertex set $\mathcal{V} = \mathbb{N}$:

1. The law of the adjacency matrix is invariant to joint permutations of its rows and columns; in other words, for any permutation $\tau \in \mathrm{Sym}(\mathcal{V})$ we have that $(a_{uv})_{u,v} \overset{d}{=} (a_{\tau(u)\tau(v)})_{u,v}$.
2. There exists a graphon W such that the law of the adjacency matrix is equivalent to a graphon model with graphon W.

The presentation above choosing the latent distribution of the vertices to be uniform is a canonical one, but can be generalized. If we instead assign latent variables $\boldsymbol{\lambda}_u \overset{\text{i.i.d}}{\sim} Q$ for some probability distribution $Q \subseteq \mathbb{R}^p$ and assign edges $a_{uv} = 1$ independently with probability $\boldsymbol{W}(\boldsymbol{\lambda}_u, \boldsymbol{\lambda}_v)$ for some symmetric function $\boldsymbol{W}$, then the law of the graph is still exchangeable, and hence equivalent to a graphon model as presented above.

One special case of a graphon model is known as a stochastic block model (SBM) [31]. The typical formulation of a SBM defines a probabilistic model on a network given a number of communities κ, a probability distribution $(\pi_i)_{i \in [\kappa]}$ on $[\kappa]$, and a symmetric matrix $P \in [0,1]^{\kappa \times \kappa}$. For $u \in \mathcal{V}$, we assign a community $C(u) \in [k]$ with probability

$$\mathbb{P}\big(C(u) = j\big) = \pi_j \text{ for } j \in [\kappa] \tag{17}$$

independently for each $u \in \mathcal{V}$. Conditional on these assignments, we then form the adjacency matrix of the network via connecting vertices independently with probability

$$\mathbb{P}\big(a_{uv} = 1 \,|\, C(u), C(v)\big) = P_{C(u),C(v)} = e_{C(u)}^T P e_{C(v)} \tag{18}$$

where $e_i \in \mathbb{R}^\kappa$ denotes the i-th unit vector in $\mathbb{R}^\kappa$. This can be defined as a graphon model as follows: forming a partition of $[0,1]$, say $(A_i)_{i \in [\kappa]}$, for which $|A_i| = \pi_i$ for $i \in [k]$, then we can define a graphon model by using latent variables $\lambda_u \overset{\text{i.i.d}}{\sim} \mathrm{Unif}[0,1]$ and a graphon function

$$W(u,v) = P_{C(u),C(v)} \text{ for } u, v \in [0,1] \qquad \text{where } C(u) = j \text{ if } u \in A_j. \tag{19}$$

The law of the above graphon model is then identical to that of the SBM defined with π and P. Such graphons are sometimes referred to as *stepfunctions*, which are graphons W which are piecewise constant on a partition $\mathcal{P} \times \mathcal{P}$ of $[0,1]^2$, where $\mathcal{P}$ is a partition of $[0,1]$.

As presented, graphon models have some shortcomings. For example, by taking expectations, if we have a graphon model on a vertex set $\mathcal{V}_n = [n]$, the average number of edges will be $n^2 \int_0^1 \int_0^1 W(x,y)\, dxdy$. This means that graphon models give rise to *dense* graphs, which is not a realistic assumption for naturally occurring networks. One way of accounting for this, particularly when considering sequences of graphs, is to consider the sequence of graphs $\mathcal{G}_n$ on $\mathcal{V}_n = [n]$ where, for each n, the generating graphon used is $W_n = \rho_n W$ where W is a graphon function, and ρ_n is a sparsifying sequence for which $\rho_n \to 0$ as $n \to \infty$. Such graphs are referred to as *sparisified graphon models*.

B Expanded derivations from Sections 1 and 2

B.1 Stochastic gradient descent and empirical risk minimization

We begin with considering the gradient updates of the form

$$\omega_u \leftarrow \omega_u - \eta \nabla_{\omega_u} \mathcal{L} \quad \text{where} \quad \mathcal{L} = \sum_{(i,j) \in \mathcal{P}} \ell_{\mathcal{P}}(\langle \omega_i, \omega_j \rangle) + \sum_{(i,j) \in \mathcal{N}} \ell_{\mathcal{N}}(\langle \omega_i, \omega_j \rangle) \tag{20}$$

performed at each iteration of stochastic gradient descent, where $\eta > 0$ is a sequence of step sizes, and $\mathcal{P}$ and $\mathcal{N}$ are random subsets of $\mathcal{V} \times \mathcal{V}$ formed at every iteration of stochastic gradient descent. Note that we can equivalently write this as

$$\mathcal{L} = \sum_{i,j} \big\{ \mathbb{1}\big[(i,j) \in \mathcal{P}\big] \ell_{\mathcal{P}}(\langle \omega_i, \omega_j \rangle) + \mathbb{1}\big[(i,j) \in \mathcal{N}\big] \ell_{\mathcal{N}}(\langle \omega_i, \omega_j \rangle) \big\}, \tag{21}$$

where we use indicator terms to allow for the summation to occur over all pairs (i, j).

Recall that stochastic gradient descent, as introduced in Robbins and Monro [55], works by the following principle: suppose we have a function of the form

$$F(\theta) := \mathbb{E}_{X \sim Q}\big[f(X, \theta)\big] \tag{22}$$

for some function $f : \mathcal{X} \times \Theta \to \mathbb{R}$ and distribution Q on $\mathcal{X}$ according to a random variable X. Moreover, suppose that we have access to an *unbiased* estimator of the gradient $\nabla_\theta F$ of F, say $g(x, \theta)$, so that $\mathbb{E}_{X \sim Q}[g(X, \theta)] = \nabla_\theta F$. One can then show in various settings [see e.g 7, 12, 13, 23, 43, 55] that the optimization scheme

$$\theta_{t+1} := \theta_t - \eta_t g(x_t; \theta) \tag{23}$$

where $x_t \overset{\text{i.i.d}}{\sim} P$ will converge to a local minima of $F(\theta)$, at least under some conditions on the step sizes $\eta_t > 0$ and the curvature of $F(\theta)$ about its local minima.

Applying this to the scenario in (21), we note that at each iteration, we sample sets $\mathcal{P} \subseteq \mathcal{V} \times \mathcal{V}$ and $\mathcal{N} \subseteq \mathcal{V} \times \mathcal{V}$ at each iteration independently across iterations, according to some probability measures $Q_{\mathcal{P}}$ and $Q_{\mathcal{N}}$ over $\mathcal{V} \times \mathcal{V}$. With these sets, we perform gradient updates as in (23)

$$\nabla_\omega \mathcal{L} = \sum_{i,j} \big\{ \mathbb{1}\big[(i,j) \in \mathcal{P}\big] \nabla_\omega \ell_{\mathcal{P}}(\langle \omega_i, \omega_j \rangle) + \mathbb{1}\big[(i,j) \in \mathcal{N}\big] \nabla_\omega \ell_{\mathcal{N}}(\langle \omega_i, \omega_j \rangle) \big\} \tag{24}$$

for each embedding vector ω, which is an unbiased estimator of $\nabla_\omega \mathcal{R}$ where

$$\mathcal{R} = \sum_{i,j} \big\{ \mathbb{P}\big((i,j) \in \mathcal{P}\big) \ell_{\mathcal{P}}(\langle \omega_i, \omega_j \rangle) + \mathbb{P}\big((i,j) \in \mathcal{N}\big) \ell_{\mathcal{N}}(\langle \omega_i, \omega_j \rangle) \big\} \tag{25}$$

as a result of the fact that e.g. $\mathbb{E}[\mathbb{1}[(i,j) \in \mathcal{P}]] = \mathbb{P}((i,j) \in \mathcal{P})$. Consequently, the procedure described in (20) attempts to minimize (25).

B.2 Embedding methods as implicit graphon learning

Write $\ell_P(y) = -\log \sigma(y)$ and $\ell_{\mathcal{N}}(y) = -\log \sigma(-y)$, and moreover suppose that $\mathcal{P}$ and $\mathcal{N}$ are randomly drawn subsets from the sets $\mathcal{E}$ and $\mathcal{V} \times \mathcal{V} \subseteq \mathcal{E}$, i.e that

$$\mathcal{P} \subseteq \{(u, v) \,:\, a_{uv} = 1\}, \qquad \mathcal{N} \subseteq \{(u, v) \,:\, a_{uv} = 0\}. \tag{26}$$

Letting $\mathcal{V} = [n]$ for some integer n, note that in the model

$$a_{uv} \,|\, \omega_u, \omega_v \sim \text{Bernoulli}\big(\sigma(\langle \omega_u, \omega_v \rangle)\big) \text{ independently for } u < v, \tag{27}$$

and setting $a_{vu} = a_{uv}$ for $v < u$, the contribution to the negative log-likelihood of a single edge (u, v) is of the form

$$\ell((u, v)) = -a_{uv} \log \sigma(\langle \omega_u, \omega_v \rangle) - (1 - a_{uv}) \log\{1 - \sigma(\langle \omega_u, \omega_v \rangle)\}. \tag{28}$$

(Recall that $\sigma(-y) = 1 - \sigma(y)$ for $y \in \mathbb{R}$.) Note that the a_{uv} are jointly independent conditional on the collection of embedding vectors. Now, if we let $\ell_{\mathcal{P}}(y) = -\log \sigma(y)$ and $\ell_{\mathcal{N}}(y) = -\log \sigma(-y)$, as $\mathcal{P}$ is a subset of $\mathcal{E}$ and $\mathcal{N}$ is a subset of $\mathcal{V} \times \mathcal{V} \setminus \mathcal{E}$, the contributions to the stochastic loss take exactly the form specified in (2), as $\ell((u, v)) = \ell_{\mathcal{P}}(\langle \omega_u, \omega_v)$ when $a_{uv} = 1$, and $\ell((u, v)) = \ell_{\mathcal{N}}(\langle \omega_u, \omega_v)$ when $a_{uv} = 0$.

B.3 Empirical risk when including regularization

We explain this using the weight decay formulation first, and then show that one has similar reasoning when using the probabilistic modelling approach. Note that when considering stochastic gradient iterations to only update individual parameters at a time, weight decay is applied per iteration to *only*

the parameters to be updated (as otherwise all of the parameters will be shrunk towards zero while waiting for the next bona-fide gradient update). Consequently, if we have a stochastic loss

$$\mathcal{L} = \sum_{(i,j)\in\mathcal{P}} \ell_{\mathcal{P}}(\langle \omega_i, \omega_j \rangle) + \sum_{(i,j)\in\mathcal{N}} \ell_{\mathcal{N}}(\langle \omega_i, \omega_j \rangle) \tag{29}$$

and we write $\mathcal{V}(\mathcal{P} \cup \mathcal{N})$ for the vertices which belong to either $\mathcal{P}$ or $\mathcal{N}$, the gradient updates for any vertex $u \in \mathcal{V}(\mathcal{P} \cup \mathcal{N})$ take the form

$$\omega_u \leftarrow \omega_u - \eta(\nabla_{\omega_u}\mathcal{L} + 2\xi\omega_u) = \omega_u - \eta\nabla_{\omega_u}\big[\mathcal{L} + \xi\|\omega_u\|_2^2\big] \tag{30}$$

and otherwise ω_u is kept as-is, meaning the gradient updates correspond to taking gradient updates with the stochastic loss

$$\sum_{u,v}\Big\{1\big[(u,v)\in\mathcal{P}\big]\ell_{\mathcal{P}}(\langle\omega_u,\omega_v\rangle) + 1\big[(u,v)\in\mathcal{N}\big]\ell_{\mathcal{N}}(\langle\omega_u,\omega_v\rangle)\Big\} + \xi\sum_u 1\big[u\in\mathcal{V}(\mathcal{P}\cup\mathcal{N})\big]. \tag{31}$$

Consequently, following the same argument as in Appendix B.1 gives the form of (8). In the probabilistic modelling formulation, we again note that the contributions to the negative log-likelihood in the subsampling regime should again only arise from vertices belonging to $\mathcal{V}(\mathcal{P} \cup \mathcal{N})$, and consequently the same argument will give the form of (8).

B.4 Simplifying the risk for certain positive/negative sampling schemes

We note that in the setting where $\mathcal{P} \subseteq \mathcal{E}_n$ and $\mathcal{N} \subseteq (\mathcal{V}_n \times \mathcal{V}_n) \setminus \mathcal{E}_n$, we can write $S(\mathcal{G}_n) = \mathcal{P} \cup \mathcal{N}$, and also

$$\ell(y, a_{ij}) = \ell_{\mathcal{P}}(y)1[a_{ij} = 1] + \ell_{\mathcal{N}}(y)1[a_{ij} = 0],$$
$$\mathbb{P}\big((i,j) \in S(\mathcal{G}_n) \,|\, \mathcal{G}_n\big) = \mathbb{P}\big((i,j) \in \mathcal{P} \,|\, \mathcal{G}_n\big)1[a_{ij} = 1] + \mathbb{P}\big((i,j) \in \mathcal{N} \,|\, \mathcal{G}_n\big)1[a_{ij} = 0],$$

and as a result, we end up obtaining (10) from (8).

C Proof of Theorem 1

We follow the style of argument given in Appendix C of [21], introducing various intermediate functions and chaining together uniform convergence bounds between these functions over sets containing the minima of both functions; consequently, we break the proof up into several parts. Note that we implicitly let the embedding dimension d depend on n throughout.

Before giving the proof, we state some results from [21] which will be used in the following proof.

Proposition 1 (Appendix C of [21])**.** *Suppose that Assumptions 2 and 3 hold. Then we have the following:*

i) (Theorem 30 and Lemma 41 of [21]) For the functions

$$\widehat{\mathcal{R}}_n(\boldsymbol{\omega}_n) := \frac{1}{n^2}\sum_{i,j\in[n],i\neq j} f_n(\lambda_i, \lambda_j, a_{ij})\ell(\langle\omega_i,\omega_j\rangle, a_{ij}),$$
$$\mathbb{E}[\widehat{\mathcal{R}}_n(\boldsymbol{\omega}_n) \,|\, \boldsymbol{\lambda}_n] := \frac{1}{n^2}\sum_{i,j\in[n],i\neq j}\sum_{x\in\{0,1\}} \tilde{f}_n(\lambda_i, \lambda_j, x)\ell(\langle\omega_i,\omega_j\rangle, x),$$

we have that

$$\sup_{\boldsymbol{\omega}_n\in([-A,A]^d)^n}\Big|\widehat{\mathcal{R}}_n(\boldsymbol{\omega}_n) - \mathbb{E}[\widehat{\mathcal{R}}_n(\boldsymbol{\omega}_n) \,|\, \boldsymbol{\lambda}_n]\Big| = O_p\Big(\frac{d^{q+1/2}}{(n\rho_n)^{1/2}}\Big).$$

ii) (Lemma 37 of [21]) For the functions

$$\mathbb{E}[\widehat{\mathcal{R}}_n^{\mathcal{P}_n}(\boldsymbol{\omega}_n) \,|\, \boldsymbol{\lambda}_n] := \frac{1}{n^2}\sum_{i,j\in[n],i\neq j}\sum_{x\in\{0,1\}} \mathcal{P}_n^{\otimes 2}[\tilde{f}_{n,x}](\lambda_i,\lambda_j)\ell(\langle\omega_i,\omega_j\rangle, x),$$
$$\mathbb{E}[\widehat{\mathcal{R}}_{n,(1)}^{\mathcal{P}_n}(\boldsymbol{\omega}_n) \,|\, \boldsymbol{\lambda}_n] := \frac{1}{n^2}\sum_{i,j\in[n]}\sum_{x\in\{0,1\}} \mathcal{P}_n^{\otimes 2}[\tilde{f}_{n,x}](\lambda_i,\lambda_j)\ell(\langle\omega_i,\omega_j\rangle, x)$$

where $\mathcal{P}_n^{\otimes 2}[\cdot]$ is the stepping operator defined in Appendix C.3, we have that

$$\sup_{\boldsymbol{\omega}_n \in ([-A,A]^d)^n} \big|\mathbb{E}[\widehat{\mathcal{R}}_n^{\mathcal{P}_n}(\boldsymbol{\omega}_n) \,|\, \boldsymbol{\lambda}_n] - \mathbb{E}[\widehat{\mathcal{R}}_{n,(1)}^{\mathcal{P}_n}(\boldsymbol{\omega}_n) \,|\, \boldsymbol{\lambda}_n]\big| = O_p\Big(\frac{d^q}{n}\Big).$$

iii) (Lemma 42 and Proposition 44 of [21]) Suppose that $X \sim \text{Multinomial}(n; p)$ where $p = (p_i)_{i\in[M]}$ for some $M > 0$, where the $p_i > 0$ and $\sum_{i=1}^M p_i = 1$. Then we have that

$$\max_{l\in[M]} \big|\frac{n^{-1}X_l - p_l}{p_l}\big|, \max_{l,l'\in[M]} \big|\frac{n^{-2}X_l X_{l'} - p_l p_{l'}}{p_l p_{l'}}\big| = O_p\Big(\Big(\frac{\log M}{n \min_l p_l}\Big)^{1/2}\Big).$$

We also require the following lemmas, whose proof are deferred to Appendix C.7.

Lemma 4. *For each $n \in \mathbb{N}$, suppose we have a compact set $\Theta_n \subseteq \mathbb{R}^{p(n)}$ for some $p(n)$ with $0 \in \Theta_n$. Moreover, suppose we have non-negative random variables $c_{ijx}^{(n)}$, $\tilde{c}_{ijx}^{(n)}$ for $i, j \in [k(n)]$ and $x \in \{0, 1\}$, and $c_i^{(n)}$, $\tilde{c}_i^{(n)}$ for $i \in [\kappa(n)]$, which satisfy the conditions*

$$\max_{i,j,x} \Big|\frac{c_{ijx}^{(n)} - \tilde{c}_{ijx}^{(n)}}{c_{ijx}^{(n)}}\Big| = O_p(r_n), \quad \max_i \Big|\frac{c_i^{(n)} - \tilde{c}_i^{(n)}}{c_i^{(n)}}\Big| = O_p(r_n),$$
$$\sum_{i,j,x} c_{ijx}^{(n)} = O_p(1), \quad \sum_i c_i^{(n)} = O_p(1),$$

for some non-negative sequence $r_n \to 0$, where in the above ratios we interpret $0/0 = 1$. Define non-negative continuous functions $\ell_{ijx}^{(n)}, \ell_i^{(n)} : \Theta_n \to \mathbb{R}$ such that $\ell_{ijx}^{(n)}(0), \ell_i^{(n)}(0) \le C$ for each $i, j \in [k(n)]$, $x \in \{0, 1\}$ and $n \in \mathbb{N}$. Finally, define the functions

$$G_n(\theta) = \sum_{i,j,x} c_{ijx}^{(n)} \ell_{ijx}^{(n)}(\theta) + \sum_i c_i^{(n)} \ell_i^{(n)}(\theta), \qquad \widetilde{G}_n(\theta) = \sum_{i,j,x} \tilde{c}_{ijx}^{(n)} \ell_{ijx}^{(n)}(\theta) + \sum_i \tilde{c}_i^{(n)} \ell_i^{(n)}(\theta)$$

for $\theta \in \Theta_n$. Then there exists a sequence of non-empty random measurable sets Ψ_n such that

$$\mathbb{P}\Big(\operatorname*{arg\,min}_{\theta_n\in\Theta_n} G_n(\theta_n) \cup \operatorname*{arg\,min}_{\theta_n\in\Theta_n} \widetilde{G}_n(\theta_n) \subseteq \Psi_n\Big) \to 1, \qquad \sup_{\theta_n\in\Psi_n} \big|G_n(\theta_n) - \widetilde{G}_n(\theta_n)\big| = O_p(r_n).$$

We note that the condition that $c_{ijx}^{(n)} = (1 + O_p(r_n))\tilde{c}_{ijx}^{(n)}$ holds uniformly over all i, j, x implies that

$$\sum_{i,j,x} c_{ijx}^{(n)} = O_p(1) \implies \sum_{i,j,x} \tilde{c}_{ijx}^{(n)} = O_p(1)$$

and so it suffices for either $\sum_{ijx} c_{ijx}^{(n)} = O_p(1)$ or $\sum_{ijx} \tilde{c}_{ijx}^{(n)} = O_p(1)$ to hold, and similarly either $\sum_i c_i^{(n)} = O_p(1)$ or $\sum_i \tilde{c}_i^{(n)} = O_p(1)$.

Lemma 5. *For each $n \in \mathbb{N}$, suppose we have a compact set $\Theta_n \subseteq \mathbb{R}^{p(n)}$ for some $p(n)$ with $0 \in \Theta_n$. Moreover, suppose we have non-negative functions $a_{n,x}, \tilde{a}_{n,x} : [0, 1]^2 \to \mathbb{R}$ and $b_n, \tilde{b}_n : [0, 1] \to \mathbb{R}$ for $n \in \mathbb{N}$, $x \in \{0, 1\}$, such that*

$$\max_x \Big\|\frac{a_{n,x} - \tilde{a}_{n,x}}{a_{n,x}}\Big\|_\infty = O(r_n), \quad \Big\|\frac{b_n - \tilde{b}_n}{b_n}\Big\|_\infty = O(r_n),$$
$$\int_{[0,1]^2} a_{n,x}\, dl dl' = O(1), \quad \int_{[0,1]} b_n\, dl = O(1)$$

for some non-negative sequence $r_n \to 0$. Define non-negative continuous functions $\ell_x : \mathbb{R} \to \mathbb{R}$ for $x \in \{0, 1\}$, along with the functions

$$G_n(\eta) = \int_{[0,1]^2} \sum_x a_{n,x}(l, l')\ell_x(\langle \eta(l), \eta(l')\rangle)\, dl dl' + \int_{[0,1]} b_n(l)\|\eta(l)\|_2^2\, dl,$$
$$\widetilde{G}_n(\eta) = \int_{[0,1]^2} \sum_x \tilde{a}_{n,x}(l, l')\ell_x(\langle \eta(l), \eta(l')\rangle)\, dl dl' + \int_{[0,1]} \tilde{b}_n(l)\|\eta(l)\|_2^2\, dl$$

defined over functions $\eta : [0,1] \to \Theta_n$*. For any fixed constant* $C > 1$*, define the set*

$$\Psi_n := \big\{\eta \,:\, G_n(\eta) \le C G_n(0) \text{ or } \widetilde{G}_n(\eta) \le C\widetilde{G}_n(0)\big\}.$$

Provided there exist minima to the functionals $G_n(\eta)$ *and* $\widetilde{G}_n(\eta)$*, we have that*

$$\arg\min_\eta G_n(\eta) \cup \arg\min_\eta \widetilde{G}_n(\eta) \subseteq \Psi_n \qquad \text{and} \qquad \sup_{\eta\in\Psi_n} \big|G_n(\eta) - \widetilde{G}_n(\eta)\big| = O(r_n).$$

We now begin with the proof of Theorem 1. Throughout, we understand that an exponent p depends on the choice of loss function, with $q = 1$ for the cross-entropy loss, and $q = 2$ for the squared loss; these will then give rise to the values of p in the exponents within the theorem statement.

C.1 Replacing the sampling probabilities

To begin, let

$$\widehat{\mathcal{R}}_n(\boldsymbol{\omega}_n) := \frac{1}{n^2}\sum_{i,j\in[n],i\neq j} f_n(\lambda_i,\lambda_j,a_{ij})\ell(\langle\omega_i,\omega_j\rangle,a_{ij}),\ \widehat{\mathcal{R}}_n^{\text{reg}}(\boldsymbol{\omega}_n) := \frac{1}{n}\sum_{i\in[n]}\tilde{g}_n(\lambda_i)\|\omega_i\|_2^2. \quad (32)$$

We then note that by applying Lemma 4 with

- $\Theta^{(n)} = ([-A,A]^d)^n$, $\theta_n = (\omega_1,\dots,\omega_n)$ with $\omega_i \in [-A,A]^d$;
- $c_{ijx}^{(n)} = \mathbb{P}((i,j) \in S(\mathcal{G}_n)\,|\,\mathcal{G}_n)\cdot 1[a_{ij} = x]$ for $i \neq j$ and 0 otherwise; $\tilde{c}_{ijx}^{(n)} = n^{-2}f_n(\lambda_i,\lambda_j,x)1[a_{ij}=x]$ for $i\neq j$ and 0 otherwise (so $\sum_{ijx}\tilde{c}_{ijx}^{(n)} = O_p(1)$ by Markov's inequality and Assumption 2);
- $c_i^{(n)} = \mathbb{P}(i \in \mathcal{V}(S(\mathcal{G}_n))\,|\,\mathcal{G}_n)$, $\tilde{c}_i^{(n)} = n^{-1}\tilde{g}_n(\lambda_i)$ (so $\sum_i c_i^{(n)} = O_p(1)$ by Markov's inequality and Assumption 2);
- $\ell_{ijx} = \ell(\langle\omega_i,\omega_j\rangle,x)$, $r_n = s_n$;

and so there exists a sequence of sets $\Psi_n^{(1)}$, containing the minima of both $\mathcal{R}_n(\boldsymbol{\omega}_n) + \xi_n\mathcal{R}_n^{\text{reg}}(\boldsymbol{\omega}_n)$ and $\widehat{\mathcal{R}}_n(\boldsymbol{\omega}_n) + \xi_n\widehat{\mathcal{R}}_n^{\text{reg}}(\boldsymbol{\omega}_n)$ with asymptotic probability one, such that

$$\sup_{\boldsymbol{\omega}_n\in\Psi_n^{(1)}}\Big|\big\{\mathcal{R}_n(\boldsymbol{\omega}_n) + \xi_n\mathcal{R}_n^{\text{reg}}(\boldsymbol{\omega}_n)\big\} - \big\{\widehat{\mathcal{R}}_n(\boldsymbol{\omega}_n) + \xi_n\widehat{\mathcal{R}}_n^{\text{reg}}(\boldsymbol{\omega}_n)\big\}\Big| = O_p(s_n). \quad (33)$$

C.2 Averaging over the adjacency structure

We now want to work with the version of the loss averaged over the realizations of the adjacency matrix of the graph $\mathcal{G}_n$, and so we introduce the function (writing $\boldsymbol{\lambda}_n = (\lambda_1,\dots,\lambda_n)$)

$$\mathbb{E}[\widehat{\mathcal{R}}_n(\boldsymbol{\omega}_n)\,|\,\boldsymbol{\lambda}_n] := \frac{1}{n^2}\sum_{i,j\in[n],i\neq j}\sum_{x\in\{0,1\}}\tilde{f}_n(\lambda_i,\lambda_j,x)\ell(\langle\omega_i,\omega_j\rangle,x). \quad (34)$$

By Proposition 1a), we have that

$$\begin{aligned}\sup_{\boldsymbol{\omega}_n\in([-A,A]^d)^n}\Big|&\big\{\widehat{\mathcal{R}}_n(\boldsymbol{\omega}_n) + \xi_n\widehat{\mathcal{R}}_n^{\text{reg}}(\boldsymbol{\omega}_n)\big\}\\ &- \big\{\mathbb{E}[\widehat{\mathcal{R}}_n(\boldsymbol{\omega}_n)\,|\,\boldsymbol{\lambda}_n] + \xi_n\widehat{\mathcal{R}}_n^{\text{reg}}(\boldsymbol{\omega}_n)\big\}\Big| = O_p\Big(\frac{d^{q+1/2}}{(n\rho_n)^{1/2}}\Big).\end{aligned} \quad (35)$$

Remark 3. *This remark can be skipped on a first reading of the theorem proof. Here, we discuss how we can obtain tighter bounds when imposing the additional constraint*

$$B_{n,d}^\infty(A_2) := \{\boldsymbol{\omega}_n \in ((\mathbb{R}^d)^n) \,:\, \Omega_{ij} = B(\omega_i,\omega_j) \text{ satisfies } \|\Omega\|_\infty \le A_2\}$$

to the domain of optimization of the embedding vectors $\boldsymbol{\omega}_n$ *is imposed. This is particularly natural when considering the squared loss, which corresponds to optimizing the risk when averaging* $(a_{ij} - \langle\omega_i,\omega_j\rangle)^2$ *over all pairs* (i,j)*; as a graphon is bounded in* $[0,1]$*, there is no need for* $\langle\omega_i,\omega_j\rangle$ *to be*

outside of the range $[0,1]$ either. With the understanding that in this remark, we write $\Omega_{ij} = \langle \omega_i, \omega_j \rangle$ for the gram matrix of the embedding vectors, we define the sets

$$Z_{n,d}(A_1) := \{\Omega \in \mathbb{R}^{n\times n} \,:\, \Omega_{ij} = \langle \omega_i, \omega_j\rangle, \|\omega_i\|_\infty \leq A_1\},$$
$$Z_n^\infty(A_2) := \{\Omega \in \mathbb{R}^{n\times n} \,:\, \max_{i,j} |\Omega_{i,j}| \leq A_2\}$$

for the constraint set placed directly on the induced matrix Ω.

We now highlight that in the proof of Theorem 30 of [21] (from which the bound just prior to the remark follows from), one looks to bound the variance term

$$v(\boldsymbol{A}_n \,|\, \boldsymbol{\lambda}_n) \leq \frac{1}{n^2}\Bigg\{ \frac{1}{n^2}\sum_{i\neq j} f_n(\lambda_i,\lambda_j,a_{ij})^2\big(\ell(\Omega_{ij},a_{ij}) - \ell(\widetilde{\Omega}_{ij},a_{ij})\big)^2 \\ + \frac{1}{n^2}\sum_{i\neq j}\mathbb{E}\Big[f_n(\lambda_i,\lambda_j,a_{ij})^2\big(\ell(\Omega_{ij},a_{ij}) - \ell(\widetilde{\Omega}_{ij},a_{ij})\big)^2 \,|\, \boldsymbol{\lambda}_n\Big]\Bigg\}$$

by some metric distance between Ω and $\widetilde{\Omega}$. To proceed, we use the alternative bound

$$\begin{aligned} v(\boldsymbol{A}_n \,|\, \boldsymbol{\lambda}_n) &\leq \frac{1}{n^2}\Bigg\{ \frac{1}{n^2}\sum_{i\neq j} f_n(\lambda_i,\lambda_j,a_{ij})^2\big(\ell(\Omega_{ij},a_{ij}) - \ell(\widetilde{\Omega}_{ij},a_{ij})\big)^2 \\ &\qquad + \frac{1}{n^2}\sum_{i\neq j}\mathbb{E}\Big[f_n(\lambda_i,\lambda_j,a_{ij})^2\big(\ell(\Omega_{ij},a_{ij}) - \ell(\widetilde{\Omega}_{ij},a_{ij})\big)^2 \,|\, \boldsymbol{\lambda}_n\Big]\Bigg\} \\ &\leq \frac{2}{n^4}\Bigg\{\max_{i,j} f_n(\lambda_i,\lambda_j,a_{ij})^2\Bigg\} \cdot L_\ell \max_{i,j}\{|\Omega_{ij}|,|\widetilde{\Omega}_{ij}|\}^{p-1}\sum_{i,j}\big(\Omega_{ij}-\widetilde{\Omega}_{ij}\big)^2 \\ &\leq \frac{2L_\ell A_2^{q-1}}{n^4}\Bigg\{\max_{i,j,x} f_n(\lambda_i,\lambda_j,x)^2\Bigg\} \cdot \|\Omega - \widetilde{\Omega}\|_F^2 \end{aligned}$$

where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix, and we note that for the cross-entropy loss (with $q=1$) and the squared loss (with $q=2$) we can write

$$|\ell(y,x) - \ell(y',x)| \leq L_\ell \max\{|y|,|y'|\}^{q-1}|y-y'|$$

for a Lipschitz constant L_ℓ. We now note that we can contain $Z_n^\infty(A_2) \cap Z_{n,d}(A_1)$ within the set

$$Z_{n,d}^F(nA_2) := \big\{\Omega \in \mathbb{R}^{n\times n} \,:\, \Omega \text{ is of rank} \leq d,\, \|\Omega\|_F \leq nA_2\big\}$$

We note that with respect to the Frobenius norm, this set has covering number

$$N(Z_{n,d}^F(nA_2), \|\cdot\|_F, \epsilon) \leq \Big(\frac{CnA_2}{\epsilon}\Big)^{2nd}$$

for some absolute constant $C>0$, and therefore by a similar argument to Lemma 41 in [21], it will be possible to conclude that $\gamma_2(Z_{n,d}^F(nA_2), \|\cdot\|_F) \leq C' n^{3/2} d^{1/2}$ for some constant C' depending on A_1 and A_2, which can then be plugged into the bound given in Theorem 30 of [21]. For the sampling schemes we consider, $\max_{i,j,x} f_n(\lambda_i,\lambda_j,x) = O(\rho_n^{-2})$, and consequently the bound we obtain is of the order $(d/n\rho_n^2)^{1/2}$, rather than $(d^3/n\rho_n)^{1/2}$. This bound is particularly effective in the non-sparse regime; in the sparse regime, one would hope for a bound of the form $(d/n\rho_n)^{1/2}$, but we are unaware as to whether such a bound is achievable.

C.3 Using a SBM approximation

We begin by working in the scenario where Assumption 3b) holds. Letting $\mathcal{P}$ be a partition of $[0,1]$ into κ parts, say $\mathcal{P} = (A_1,\ldots,A_\kappa)$, we introduce the stepping operators defined by

$$\mathcal{P}^{\otimes 2}[h](x,y) = \frac{1}{|A_l||A_{l'}|}\int_{A_l\times A_{l'}} h(x',y')\,dx'dy' \text{ if } (x,y)\in A_l\times A_{l'},$$
$$\mathcal{P}[h](x) = \frac{1}{|A_l|}\int_{A_l} h(x')\,dx' \text{ if } x \in A_l$$

for any symmetric measurable function $h : [0,1]^2 \to \mathbb{R}$ and measurable function $h : [0,1] \to \mathbb{R}$ respectively. With this, let $\mathcal{P}_n = (A_{n1}, \ldots, A_{n\kappa(n)})$ be a sequence of partitions containing $\kappa(n)$ intervals of size $|A_{nl}| \asymp n^{-\alpha}$ for some constant $\alpha > 0$, and then introduce the functions

$$\mathbb{E}[\widehat{\mathcal{R}}_n^{\mathcal{P}_n}(\boldsymbol{\omega}_n) \,|\, \boldsymbol{\lambda}_n] := \frac{1}{n^2} \sum_{i,j \in [n], i \neq j} \sum_{x \in \{0,1\}} \mathcal{P}_n^{\otimes 2}[\tilde{f}_{n,x}](\lambda_i, \lambda_j) \ell(\langle \omega_i, \omega_j \rangle, x), \tag{36}$$

$$\widehat{\mathcal{R}}_n^{\mathrm{reg},\mathcal{P}_n}(\boldsymbol{\omega}_n) := \frac{1}{n} \sum_{i \in [n]} \mathcal{P}_n[\tilde{g}_n](\lambda_i) \|\omega_i\|_2^2, \tag{37}$$

where we make the abbreviation $\tilde{f}_{n,x}(\lambda_i, \lambda_j) := \tilde{f}_n(\lambda_i, \lambda_j, x)$. We note that as $\tilde{f}_{n,x}$ and $\tilde{g}_n$ are uniformly bounded away from zero by M^{-1}, and because they are Hölder of exponent β, we can apply Lemma C.6 of [65] to obtain that

$$\left\| \frac{\tilde{f}_{n,x} - \mathcal{P}_n^{\otimes 2}[\tilde{f}_{n,x}]}{\tilde{f}_{n,x}} \right\|_\infty = O(n^{-\alpha\beta}), \qquad \left\| \frac{\tilde{g}_n - \mathcal{P}_n[\tilde{g}_n]}{\tilde{g}_n} \right\|_\infty = O(n^{-\alpha\beta}). \tag{38}$$

This, along with the uniform boundedness conditions on the $\tilde{f}_{n,x}$ and $\tilde{g}_n$ given in Assumption 3, allow us to apply Lemma 4 to find that there exists a sequence of sets $\Psi_n^{(2)}$ for which the minima of both $\mathbb{E}[\widehat{\mathcal{R}}_n(\boldsymbol{\omega}_n) \,|\, \boldsymbol{\lambda}_n] + \xi_n \widehat{\mathcal{R}}_n^{\mathrm{reg}}(\boldsymbol{\omega}_n)$ and $\mathbb{E}[\widehat{\mathcal{R}}_n^{\mathcal{P}_n}(\boldsymbol{\omega}_n) \,|\, \boldsymbol{\lambda}_n] + \xi_n \widehat{\mathcal{R}}_n^{\mathrm{reg},\mathcal{P}_n}(\boldsymbol{\omega}_n)$ are contained within it with asymptotic probability 1, and

$$\begin{aligned} \sup_{\boldsymbol{\omega}_n \in \Psi_n^{(2)}} \Big| &\big\{ \mathbb{E}[\widehat{\mathcal{R}}_n(\boldsymbol{\omega}_n) \,|\, \boldsymbol{\lambda}_n] + \xi_n \widehat{\mathcal{R}}_n^{\mathrm{reg}}(\boldsymbol{\omega}_n) \big\} \\ &- \big\{ \mathbb{E}[\widehat{\mathcal{R}}_n^{\mathcal{P}_n}(\boldsymbol{\omega}_n) \,|\, \boldsymbol{\lambda}_n] + \xi_n \widehat{\mathcal{R}}_n^{\mathrm{reg},\mathcal{P}_n}(\boldsymbol{\omega}_n) \big\} \Big| = O_p(n^{-\alpha\beta}). \end{aligned} \tag{39}$$

Note that in the case where Assumption 3a) holds, this step is not necessary, and so we can take the above bound to be equal to zero.

C.4 Adding the contribution of the diagonal term

We note that in the definition of $\mathbb{E}[\widehat{\mathcal{R}}_n^{\mathcal{P}_n}(\boldsymbol{\omega}_n) \,|\, \boldsymbol{\lambda}_n]$, the summation does not include any $i = j$ terms; if we introduce

$$\mathbb{E}[\widehat{\mathcal{R}}_{n,(1)}^{\mathcal{P}_n}(\boldsymbol{\omega}_n) \,|\, \boldsymbol{\lambda}_n] := \frac{1}{n^2} \sum_{i,j \in [n]} \sum_{x \in \{0,1\}} \mathcal{P}_n^{\otimes 2}[\tilde{f}_{n,x}](\lambda_i, \lambda_j) \ell(\langle \omega_i, \omega_j \rangle, x), \tag{40}$$

then by Proposition 1b), we have that

$$\begin{aligned} \sup_{\boldsymbol{\omega}_n \in ([-A,A]^d)^n} \Big| &\{ \mathbb{E}[\widehat{\mathcal{R}}_n^{\mathcal{P}_n}(\boldsymbol{\omega}_n) \,|\, \boldsymbol{\lambda}_n] + \xi_n \widehat{\mathcal{R}}_n^{\mathrm{reg},\mathcal{P}_n}(\boldsymbol{\omega}_n) \} \\ &- \{ \mathbb{E}[\widehat{\mathcal{R}}_{n,(1)}^{\mathcal{P}_n}(\boldsymbol{\omega}_n) \,|\, \boldsymbol{\lambda}_n] + \xi_n \widehat{\mathcal{R}}_n^{\mathrm{reg},\mathcal{P}_n}(\boldsymbol{\omega}_n) \} \Big| = O_p\Big(\frac{d^q}{n}\Big). \end{aligned} \tag{41}$$

C.5 Linking minimizing embeddings to minimizing kernels

We now want to reason about the minima of the function $\mathbb{E}[\widehat{\mathcal{R}}_{n,(1)}^{\mathcal{P}_n}(\boldsymbol{\omega}_n) \,|\, \boldsymbol{\lambda}_n] + \xi_n \widehat{\mathcal{R}}_n^{\mathrm{reg},\mathcal{P}_n}(\boldsymbol{\omega}_n)$. We denote

$$p_n(l) := |A_{nl}|, \qquad \mathcal{A}_n(l) := \{i \in [n] \,:\, \lambda_i \in A_{nl}\}, \qquad \widehat{p}_n(l) := n^{-1} |\mathcal{A}_n(l)|,$$
$$c_{f,n}(l, l', x) := \frac{1}{p_n(l) p_n(l')} \int_{A_{nl} \times A_{nl'}} \tilde{f}_n(\lambda, \lambda', x) \, d\lambda d\lambda', \qquad c_{g,n}(l) := \frac{1}{p_n(l)} \int_{A_{nl}} \tilde{g}_n(\lambda) \, d\lambda.$$

Consider writing

$$\tilde{\omega}_i = \frac{1}{|\mathcal{A}_n(l)|} \sum_{j \in \mathcal{A}_n(l)} \omega_j \text{ if } i \in \mathcal{A}_n(l), \qquad \tilde{\boldsymbol{\omega}}_n = (\tilde{\omega}_i)_{i \in [n]}, \tag{42}$$

given any set of embedding vectors $\boldsymbol{\omega}_n$. As $\ell(y,x)$ is strictly convex in $y \in \mathbb{R}$ for $x \in \{0,1\}$ and $\|\cdot\|_2^2$ is also strictly convex, by Jensen's inequality we have that

$$\begin{aligned}
&\mathbb{E}[\widehat{\mathcal{R}}_{n,(1)}^{\mathcal{P}_n}(\boldsymbol{\omega}_n) \,|\, \boldsymbol{\lambda}_n] + \xi_n \widehat{\mathcal{R}}_n^{\text{reg},\mathcal{P}_n}(\boldsymbol{\omega}_n) \\
&\quad= \sum_{l,l' \in [\kappa(n)]} \widehat{p}_n(l)\widehat{p}_n(l') \sum_x \Big\{ \frac{c_{f,n}(l,l',x)}{|\mathcal{A}_n(l)||\mathcal{A}_n(l')|} \sum_{\substack{i \in \mathcal{A}_n(l) \\ j \in \mathcal{A}_n(l')}} \ell(\langle \omega_i, \omega_j \rangle, x) \Big\} \\
&\qquad\qquad + \sum_{l \in [\kappa(n)]} \widehat{p}_n(l) \frac{c_{g,n}(l)}{|\mathcal{A}_n(l)|} \sum_{i \in \mathcal{A}_n(l)} \|\omega_i\|_2^2 \\
&\quad\geq \sum_{l,l' \in [\kappa(n)]} \widehat{p}_n(l)\widehat{p}_n(l') \sum_x c_{f,n}(l,l',x) \ell\Big(\sum_{\substack{i \in \mathcal{A}_n(l) \\ j \in \mathcal{A}_n(l')}} \langle \omega_i, \omega_j \rangle, x \Big) \\
&\qquad\qquad + \sum_{l \in [\kappa(n)]} \widehat{p}_n(l) c_{g,n}(l) \Big\| \frac{1}{|\mathcal{A}_n(l)|} \sum_{i \in \mathcal{A}_n(l)} \omega_i \Big\|_2^2 \\
&\quad= \sum_{l,l' \in [\kappa(n)]} \widehat{p}_n(l)\widehat{p}_n(l') \sum_x c_{f,n}(l,l',x) \ell\Big(\langle \sum_{i \in \mathcal{A}_n(l)} \frac{\omega_i}{|\mathcal{A}_n(l)|}, \sum_{j \in \mathcal{A}_n(l')} \frac{\omega_j}{|\mathcal{A}_n(l')|} \rangle, x \Big) \\
&\qquad\qquad + \sum_{l \in [\kappa(n)]} \widehat{p}_n(l) c_{g,n}(l) \Big\| \frac{1}{|\mathcal{A}_n(l)|} \sum_{i \in \mathcal{A}_n(l)} \omega_i \Big\|_2^2 \\
&\quad= \mathbb{E}[\widehat{\mathcal{R}}_{n,(1)}^{\mathcal{P}_n}(\widetilde{\boldsymbol{\omega}}_n) \,|\, \boldsymbol{\lambda}_n] + \xi_n \widehat{\mathcal{R}}_n^{\text{reg},\mathcal{P}_n}(\widetilde{\boldsymbol{\omega}}_n),
\end{aligned}$$

with equality if and only if the ω_i are equal across each of the sets $\mathcal{A}_n(l)$. In particular, this means that to minimize $\mathbb{E}[\widehat{\mathcal{R}}_{n,(1)}^{\mathcal{P}_n}(\boldsymbol{\omega}_n) \,|\, \boldsymbol{\lambda}_n] + \xi_n \widehat{\mathcal{R}}_n^{\text{reg},\mathcal{P}_n}(\boldsymbol{\omega}_n)$, if we define

$$\begin{aligned}
\widehat{I}_n^{\mathcal{P}_n}(\tilde{\omega}_1, \ldots, \tilde{\omega}_{\kappa(n)}) &:= \sum_{l,l' \in [\kappa(n)]} \widehat{p}_n(l)\widehat{p}_n(l') \sum_x c_{f,n}(l,l',x) \ell(\langle \tilde{\omega}_l, \tilde{\omega}_{l'} \rangle, x) \\
\widehat{I}_n^{\text{reg},\mathcal{P}_n}(\tilde{\omega}_1, \ldots, \tilde{\omega}_{\kappa(n)}) &:= \sum_{l \in [\kappa(n)]} \widehat{p}_n(l) c_{g,n}(l) \|\tilde{\omega}_l\|_2^2,
\end{aligned}$$

then it suffices to minimize $\widehat{I}_n^{\mathcal{P}_n}(\tilde{\omega}_1, \ldots, \tilde{\omega}_{\kappa(n)}) + \xi_n \widehat{I}_n^{\text{reg},\mathcal{P}_n}(\tilde{\omega}_1, \ldots, \tilde{\omega}_{\kappa(n)})$, as the $\tilde{\omega}_i$ are constant across $i \in \mathcal{A}_n(l)$. In other words, the above argument has just showed that

$$\begin{aligned}
&\min_{\boldsymbol{\omega}_n \in ([-A,A]^d)^n} \big\{ \mathbb{E}[\widehat{\mathcal{R}}_{n,(1)}^{\mathcal{P}_n}(\boldsymbol{\omega}_n) \,|\, \boldsymbol{\lambda}_n] + \xi_n \widehat{\mathcal{R}}_n^{\text{reg},\mathcal{P}_n}(\boldsymbol{\omega}_n) \big\} \\
&\qquad = \min_{(\tilde{\omega}_i) \in ([-A,A]^d)^{\kappa(n)}} \big\{ \widehat{I}_n^{\mathcal{P}_n}(\tilde{\omega}_1, \ldots, \tilde{\omega}_{\kappa(n)}) + \xi_n \widehat{I}_n^{\text{reg},\mathcal{P}_n}(\tilde{\omega}_1, \ldots, \tilde{\omega}_{\kappa(n)}) \big\}.
\end{aligned} \tag{43}$$

We note that $\widehat{I}_n^{\mathcal{P}_n}$ and $\widehat{I}_n^{\text{reg},\mathcal{P}_n}$ are stochastic, as they depend on the random variables $\widehat{p}_n(l)$. To remove the stochasticity, we introduce the functions

$$\begin{aligned}
I_n^{\mathcal{P}_n}(\tilde{\omega}_1, \ldots, \tilde{\omega}_{\kappa(n)}) &:= \sum_{l,l' \in [\kappa(n)]} p_n(l) p_n(l') \sum_x c_{f,n}(l,l',x) \ell(\langle \tilde{\omega}_l, \tilde{\omega}_{l'} \rangle, x) \\
I_n^{\text{reg},\mathcal{P}_n}(\tilde{\omega}_1, \ldots, \tilde{\omega}_{\kappa(n)}) &:= \sum_{l \in [\kappa(n)]} p_n(l) c_{g,n}(l) \|\tilde{\omega}_l\|_2^2.
\end{aligned}$$

As by Proposition 1c) we have that

$$\max_{l,l' \in [\kappa(n)]} \Big| \frac{\widehat{p}_n(l)\widehat{p}_n(l') - p_n(l)p_n(l')}{p_n(l)p_n(l')} \Big| = \begin{cases} O_p\Big(\Big(\frac{\log \kappa}{n} \Big)^{1/2} \Big) & \text{under Assumption 3a)} \\ O_p\Big(\frac{\sqrt{\log n}}{n^{1/2-\alpha/2}} \Big) & \text{under Assumption 3b)} \end{cases}$$

$$\max_{l \in [\kappa(n)]} \Big| \frac{\widehat{p}_n(l) - p_n(l)}{p_n(l)} \Big| = \begin{cases} O_p\Big(\Big(\frac{\log \kappa}{n} \Big)^{1/2} \Big) & \text{under Assumption 3a)} \\ O_p\Big(\frac{\sqrt{\log n}}{n^{1/2-\alpha/2}} \Big) & \text{under Assumption 3b)} \end{cases}$$

and moreover that the $\widehat{p}_n(l)$ and $p_n(l)$ sum to 1, we can apply Lemma 4 to argue that there exists a sequence of sets $\Psi_n^{(3)}$ which contains both the minima of $\widehat{I}_n^{\mathcal{P}_n}((\tilde{\omega}_i)_{i=1}^{\kappa(n)}) + \xi_n \widehat{I}_n^{\text{reg},\mathcal{P}_n}((\tilde{\omega}_i)_{i=1}^{\kappa(n)})$ and $I_n^{\mathcal{P}_n}((\tilde{\omega}_i)_{i=1}^{\kappa(n)}) + \xi_n I_n^{\text{reg},\mathcal{P}_n}((\tilde{\omega}_i)_{i=1}^{\kappa(n)})$ with asymptotic probability 1, and that

$$\sup_{(\tilde{\omega}_i)_i \in \Psi_n^{(3)}} \left| \left\{ \widehat{I}_n^{\mathcal{P}_n}((\tilde{\omega}_i)_{i=1}^{\kappa(n)}) + \xi_n \widehat{I}_n^{\text{reg},\mathcal{P}_n}((\tilde{\omega}_i)_{i=1}^{\kappa(n)}) \right\} - \left\{ I_n^{\mathcal{P}_n}((\tilde{\omega}_i)_{i=1}^{\kappa(n)}) + \xi_n I_n^{\text{reg},\mathcal{P}_n}((\tilde{\omega}_i)_{i=1}^{\kappa(n)}) \right\} \right| = \begin{cases} O_p\Big(\Big(\frac{\log \kappa}{n}\Big)^{1/2}\Big) & \text{under Assumption 3a)} \\ O_p\Big(\frac{\sqrt{\log n}}{n^{1/2-\alpha/2}}\Big) & \text{under Assumption 3b)} \end{cases} \tag{44}$$

To transition from embedding vectors to kernels $K(l,l') = \langle \eta(l), \eta(l') \rangle$ for $\eta : [0,1] \to [-A,A]^d$, we note that as we can write

$$\mathcal{I}_n^{\mathcal{P}_n}[K] = \sum_{l,l' \in [\kappa(n)]} p_n(l) p_n(l') \sum_x \frac{c_{f,n}(l,l',x)}{p_n(l)p_n(l')} \int_{A_{nl} \times A_{nl'}} \ell(\langle \eta(\lambda), \eta(\lambda') \rangle, x) \, d\lambda d\lambda',$$

$$\mathcal{I}_n^{\text{reg},\mathcal{P}_n}[K] = \sum_{l \in [\kappa(n)]} p_n(l) \frac{c_{g,n}(l)}{p_n(l)} \int_{A_{nl}} \|\eta(\lambda)\|_2^2 \, d\lambda,$$

by the same Jensen's inequality argument used to obtain (43), we get that

$$\min_{(\tilde{\omega}_i)_i \in ([-A,A]^d)^{\kappa(n)}} \left\{ I_n^{\mathcal{P}_n}((\tilde{\omega}_i)_{i=1}^{\kappa(n)}) + \xi_n I_n^{\text{reg},\mathcal{P}_n}((\tilde{\omega}_i)_{i=1}^{\kappa(n)}) \right\} = \min_{K \in \mathcal{Z}_d^{\geq 0}(A)} \left\{ \mathcal{I}_n^{\mathcal{P}_n}[K] + \xi_n \mathcal{I}_n^{\text{reg},\mathcal{P}_n}[K] \right\}, \tag{45}$$

where the correspondence between the minimizing $K(l,l') = \langle \eta(l), \eta(l') \rangle$ and $\tilde{\omega}_i$ is given by $\eta(l) = \tilde{\omega}_i$ for $i \in A_{nl}$.

The final step is to remove the approximation terms $\mathcal{P}_n^{\otimes 2}[\tilde{f}_{n,x}]$ and $\mathcal{P}_n[\tilde{g}_n]$ from $\mathcal{I}_n^{\mathcal{P}_n}[K]$ and $\mathcal{I}_n^{\text{reg},\mathcal{P}_n}[K]$ in order to get to $\mathcal{I}_n[K]$ and $\mathcal{I}_n^{\text{reg}}[K]$. To do so, we can use (38) and Lemma 5 to obtain that there exists a set $\Psi_n^{(4)}$ containing both the minima of $\mathcal{I}_n[K] + \xi_n \mathcal{I}_n^{\text{reg}}[K]$ and $\mathcal{I}_n^{\mathcal{P}_n}[K] + \xi_n \mathcal{I}_n^{\text{reg},\mathcal{P}_n}[K]$ (which exist by Proposition 2) and

$$\sup_{K \in \Psi_n^{(4)}} \left| \left\{ \mathcal{I}_n^{\mathcal{P}_n}[K] + \xi_n \mathcal{I}_n^{\text{reg},\mathcal{P}_n}[K] \right\} - \left\{ \mathcal{I}_n[K] + \xi_n \mathcal{I}_n^{\text{reg}}[K] \right\} \right| = O(n^{-\alpha\beta}). \tag{46}$$

C.6 Combining to obtain rates of convergence

To conclude, we first note that given uniform convergence bounds of two functions on a set containing both of their minima, we can argue convergence of their minimal values; indeed if a set A contains minima x_f and x_g to some functions g, then

$$\min_x f(x) - \min_x g(x) = \min_x f(x) - g(x_g) \leq f(x_g) - g(x_g) \leq \sup_{x \in A} |f(x) - g(x)|,$$

and similarly so for $\min_x g(x) - \min_x f(x)$. (We note that Proposition 2 argues that all the relevant infimal values of the minimizers of the $\mathcal{I}_n[K] + \xi_n \mathcal{I}_n^{\text{reg}}[K]$ are attained.) Therefore, using this fact and chaining together the bounds in (33), (35), (39), (41), (43), (44) and (46), we get when Assumption 3b) holds that

$$\left| \min_{\boldsymbol{\omega}_n \in ([-A,A]^d)^n} \left\{ \mathcal{R}_n(\boldsymbol{\omega}_n) + \xi_n \mathcal{R}_n^{\text{reg}}(\boldsymbol{\omega}_n) \right\} - \min_{K \in \mathcal{Z}_d^{\geq 0}(A)} \left\{ \mathcal{I}_n[K] + \xi_n \mathcal{I}_n^{\text{reg}}[K] \right\} \right| = O_p\Big(s_n + \frac{d^{3/2} \rho_n^{-1/2}}{n^{1/2}} + n^{-\alpha\beta} + \frac{\sqrt{\log n}}{n^{1/2-\alpha/2}} \Big). \tag{47}$$

(We note that the $O_p(d^q/n)$ term from (41) is negligible.) To conclude, we simply pick an optimal choice of α, which we take to be $\alpha = 1/(1+2\beta)$, which gives the stated bound. In the case where Assumption 3a) holds, the term from the SBM approximation disappears and the $\sqrt{\log n}/n^{1/2-\alpha/2}$ term becomes $(\log \kappa/n)^{1/2}$, giving the stated bound in this regime.

C.7 Proofs of useful lemmata

Proof of Lemma 4. We begin by noting that as each of the $G_n(\theta)$ and $\widetilde{G}_n(\theta)$ are continuous functions defined on compact sets, the minima sets of each of the functions are non-empty. We now define the sets

$$\Psi_n := \left\{ \theta_n \in \Theta^{(n)} \,:\, G_n(\theta_n) \leq 2C \sum_{i,j,x} c_{ijx}^{(n)} + 2C \sum_i c_i^{(n)} \right\},$$
$$\widetilde{\Psi}_n := \left\{ \theta_n \in \Theta^{(n)} \,:\, \widetilde{G}_n(\theta_n) \leq C \sum_{i,j,x} \tilde{c}_{ijx}^{(n)} + C \sum_i \tilde{c}_i^{(n)} \right\},$$

and note that $0 \in \Psi_n$, $0 \in \widetilde{\Psi}_n$ for each n, and therefore we also have that $\arg\min_{\theta_n \in \Theta^{(n)}} G_n(\theta) \subseteq \Psi_n$ and $\arg\min_{\theta_n \in \Theta^{(n)}} \widetilde{G}_n(\theta) \subseteq \widetilde{\Psi}_n$. We now want to argue that $\mathbb{P}(\widetilde{\Psi}_n \subseteq \Psi_n) \to 1$ as $n \to \infty$. Note that for any $\theta_n \in \widetilde{\Psi}_n$, we have that

$$\begin{aligned} G_n(\theta_n) &= \sum_{i,j,x} \frac{c_{ijx}^{(n)}}{\tilde{c}_{ijx}^{(n)}} \tilde{c}_{ijx}^{(n)} \ell_{ijx}^{(n)}(\theta) + \sum_i \frac{c_i^{(n)}}{\tilde{c}_i^{(n)}} \tilde{c}_i^{(n)} \ell_i^{(n)}(\theta) \\ &\leq (1 + O_p(r_n)) \widetilde{G}_n(\theta_n) \leq C(1 + O_p(r_n)) \Big\{ \sum_{i,j,x} \tilde{c}_{ijx}^{(n)} + \sum_i \tilde{c}_i^{(n)} \Big\}. \end{aligned}$$

As by Cauchy's third inequality we have that

$$\frac{\sum_{i,j,x} \tilde{c}_{ijx}^{(n)}}{\sum_{i,j,x} c_{ijx}^{(n)}} \leq \max_{i,j,x} \frac{\tilde{c}_{ijx}^{(n)}}{c_{ijx}^{(n)}} = 1 + O_p(r_n),$$

and similarly $\sum_i \tilde{c}_i^{(n)} \leq (1 + O_p(r_n)) \sum_i c_i^{(n)}$, it follows that

$$G_n(\theta_n) \leq C(1 + O_p(r_n)) \Big\{ \sum_{i,j,x} \tilde{c}_{ijx}^{(n)} + \sum_i \tilde{c}_i^{(n)} \Big\} \overset{w.h.p}{\leq} 2C \Big\{ \sum_{i,j,x} c_{ijx}^{(n)} + \sum_i c_i^{(n)} \Big\}$$

once n is sufficiently large, and therefore $\theta_n \in \widetilde{\Psi}_n$ for n sufficiently large. In particular, as the above argument holds freely of the choice of θ_n, we have that $\widetilde{\Psi}_n \subseteq \Psi_n$ with asymptotic probability one. With this, we now note that $\sup_{\theta_n \in \Psi_n} G_n(\theta_n) = O_p(1)$ (due to the condition on the sum of the $c_{ijx}^{(n)}$ and $c_i^{(n)}$), and consequently we have that for all $\theta_n \in \Psi_n$

$$\begin{aligned} \big| G_n(\theta_n) - \widetilde{G}_n(\theta_n) \big| &\leq \max_{i,j,x} \Big| \frac{c_{ijx}^{(n)} - \tilde{c}_{ijx}^{(n)}}{c_{ijx}^{(n)}} \Big| \cdot \sum_{i,j,x} c_{ijx}^{(n)} \ell_{ijx}^{(n)}(\theta) + \max_i \Big| \frac{c_i^{(n)} - \tilde{c}_i^{(n)}}{c_i^{(n)}} \Big| \cdot \sum_i c_i^{(n)} \ell_i^{(n)}(\theta) \\ &\leq O_p(r_n) G_n(\theta_n) \leq O_p(r_n) \end{aligned}$$

with the bound holding uniformly over the choice of θ_n, giving the stated conclusion. □

Proof of Lemma 5. The proof follows the exact same style of argument as in Lemma 4, so we skip repeating the details. □

D Proof of Theorem 2

Before proving any results, we introduce some useful facts from functional analysis; the terminology and basic properties used below can be found in standard references such as e.g. [5, 14, 49]. Throughout, we will write μ_n to refer to the measure $\mu_n(A) := \int_A \tilde{g}_n d\mu$, define for all Borel sets of $[0, 1]$, where μ is the regular Lebesgue measure on $[0, 1]$, and write e.g. $L^2([0, 1], \mu_n)$ or $L^2(\mu_n)$ for the associated Lebesgue space of square integrable random variables. We note that as it assumed that the $\tilde{g}_n$ are uniformly bounded away from zero and uniformly bounded above by Assumption 3,

$h \in L^2(\mu_n)$ iff $h \in L^2(\mu)$. For any function $K \in L^2([0,1]^2, \mu_n^{\otimes 2})$ (where we write $\mu_n^{\otimes 2}$ for the product measure of μ_n with itself), we introduce the associated operator

$$T_K : L^2(\mu_n) \to L^2(\mu_n), \qquad T_K[f](x) = \int_0^1 K(x,y) f(y) d\mu_n(y). \tag{48}$$

The above operator is Hilbert-Schmidt, where all Hilbert-Schmidt operators $L^2(\mu_n) \to L^2(\mu_n)$ can be written in the above form for some kernel $K \in L^2([0,1]^2, \mu_n^{\otimes 2})$; moreover T_K is self-adjoint (so $T_K^* = T_K$) iff K is symmetric. The above identification corresponds to an isometric isomorphism between the Hilbert spaces $L^2(\mu_n)$ and the Hilbert-Schmidt operators, via [e.g 29, Theorem 8.4.8] the formula

$$\mathrm{Tr}(T_K^* T_L) = \langle K, L \rangle_{L^2([0,1]^2, \mu_n^{\otimes 2})} = \int_{[0,1]^2} \overline{K(y,x)} L(x,y)\, d\mu_n(x) d\mu_n(y), \tag{49}$$

which gives rise to the corresponding norm formula $\|T_K\|_{HS}^2 = \|K\|_{L^2(\mu_n)}$. Writing $\mathcal{S}(L^2(\mu_n))$ for the space of linear operators $L^2(\mu_n) \to L^2(\mu_n)$ with finite trace or nuclear norm $\|T\|_1 < \infty$ (referred to as the space of trace class operators), $\mathcal{K}(L^2(\mu_n))$ for the space of compact linear operators $L^2(\mu_n) \to L^2(\mu_n)$, and $\mathcal{B}(L^2(\mu_n))$ for the space of bounded linear operators $L^2(\mu_n) \to L^2(\mu_n)$ with norm $\|\cdot\|_{\mathrm{op}}$, we have that [e.g 57, Theorem 3.3.9]

- $\mathcal{S}(L^2(\mu_n)) \cong (\mathcal{K}(L^2(\mu_n)))^*$ via the mapping $T \in \mathcal{S}(L^2(\mu_n)) \mapsto [A \mapsto \mathrm{Tr}(AT)]$;
- $\mathcal{B}(L^2(\mu_n)) \cong (\mathcal{S}(L^2(\mu_n)))^*$ via the mapping $A \in \mathcal{B}(L^2(\mu_n)) \mapsto [T \mapsto \mathrm{Tr}(AT)]$.

Consequently, this allows us to argue that the trace norm $\|\cdot\|_1$ is weak* lower semi-continuous on $\mathcal{S}(L^2(\mu_n))$, and that its closed level sets are weak* compact by Banach-Alaoglu. We also note that we have the inclusions

$$\{\text{finite rank}\} \subset \{\text{trace class}\} \subset \{\text{Hilbert-Schmidt}\} \subset \{\text{compact operators}\} \subset \{\text{bounded operators}\}.$$

Operators which satisify $\langle T_K[f], f \rangle \geq 0$ for all $f \in L^2(\mu_n)$ are called positive operators[1]; for positive operators we have that the trace norm is equal simply to the trace. With this, we now are in a position to prove the results needed to talk about minimizers of $\mathcal{I}_n[K] + \xi_n \mathcal{I}_n^{\mathrm{reg}}[K]$ over various sets of functions K.

Proposition 2. *For $K \in \mathcal{Z}_{fr}^{\geq 0}(A) := \cup_{d \geq 1} \mathcal{Z}_d^{\geq 0}(A)$, writing $K(l,l') = \sum_{i=1}^d \eta_i(l)\eta_i(l')$ for some d and functions $\eta_i : [0,1] \to [-A, A]$, we define*

$$\mathcal{I}_n[K] = \int_{[0,1]^2} \sum_{x \in \{0,1\}} \tilde{f}_n(l,l',x) \ell(K(l,l'),x)\, dl dl', \qquad \mathcal{I}_n^{reg}[K] = \int_{[0,1]} \|\eta_i(l)\|_2^2 \cdot \tilde{g}_n(l)\, dl,$$

where we recall that $\tilde{f}_n$ and $\tilde{g}_n$ are as given in Assumptions 2 and 3, and $\ell(y,x)$ is either the cross-entropy loss or the squared loss function; we introduce a variable q for which $q = 1$ applies to the cross-entropy loss, and $q = 2$ for the squared loss. Treat n as fixed. Write μ_n for the measure $\mu_n(A) := \int_A \tilde{g}_n\, d\mu$ where μ is the Lebesgue measure on $[0,1]$. Then we have the following:

i) For $K \in \mathcal{Z}_{fr}^{\geq 0}(A)$, $\mathcal{I}_n^{reg}[K] = \mathrm{Tr}[T_K]$ where

$$T_K : L^2(\mu_n) \to L^2(\mu_n), \qquad T_K[f](x) = \int_0^1 K(x,y) f(y) \tilde{g}_n(y)\, dy.$$

ii) The set $\mathcal{Z}_{fr}^{\geq 0}(A)$ is free of A, and so we can let $Z^{\geq 0}$ denote the weak closure of $\mathcal{Z}_{fr}^{\geq 0}(A)$ in $\mathcal{S}(L^2(\mu_n))$.*

iii) $\mathcal{I}_n^{reg}[K]$ extends uniquely to a weak lower semi-continuous function, namely the trace, on $\mathcal{Z}^{\geq 0}$, and to the larger domain of the positive trace-class operators on $L^2(\mu_n)$. Consequently, we write $\mathcal{I}_n^{reg}[K] = \mathrm{Tr}[T_K]$ for $K \in \mathcal{Z}^{\geq 0}$, or more generally any symmetric function K for which T_K is positive.*

[1] We note that unlike in finite dimensions, we usually do not distingush between operators which are positive definite as compared to being only non-negative definite.

iv) $\mathcal{I}_n[K]$ is finite for all symmetric functions K for which T_K is a positive operator and $\mathcal{I}_n^{reg}[K] < \infty$; $\mathcal{I}_n[K]$ is strictly convex in K; and $\mathcal{I}_n[K]$ is weak lower semi-continuous with respect to the topology on $\mathcal{S}(L^2(\mu_n))$.*

v) We have the local Lipschitz property

$$\begin{aligned}\left|\mathcal{I}_n[K] - \mathcal{I}_n[L]\right| &\leq 2M^3\Big(\|K\|_{L^2(\mu_n^{\otimes 2})} + \|L\|_{L^2(\mu_n^{\otimes 2})}\Big)^{q-1}\|K-L\|_{L^2(\mu_n^{\otimes 2})} \\ &= 2M^3\Big(\|T_K\|_{HS} + \|T_L\|_{HS}\Big)^{q-1}\|T_K - T_L\|_{HS}.\end{aligned}$$

vi) For any $\xi_n \geq 0$, we have that $\mathcal{I}_n[K] + \xi_n\mathcal{I}_n^{reg}[K]$ is a strictly convex function in K, which is weak lower semi-continuous with respect to the topology on $\mathcal{S}(L^2(\mu_n))$.*

vii) For each d, there exists at least one minimizer of $\mathcal{I}_n[K] + \xi_n\mathcal{I}_n^{reg}[K]$ over $\mathcal{Z}_d^{\geq 0}(A)$, and there exists a unique minimizer to $\mathcal{I}_n[K] + \xi_n\mathcal{I}_n^{reg}[K]$ over $\mathcal{Z}^{\geq 0}$.

viii) When Assumption 3a) holds, the minima of $\mathcal{I}_n[K] + \xi_n\mathcal{I}_n^{reg}[K]$ over $\mathcal{Z}^{\geq 0}$ can be determined via a finite dimensional convex program; write K_n^* for such a minima. Moreover, there exists some $r = r(n) \leq \kappa$ such that K_n^* is of rank $r(n)$, and moreover as soon as $d \geq r(n)$ and $A \geq (\kappa - 1)\|K_n^*\|_\infty$, we have that the minima of $\mathcal{I}_n[K] + \xi_n\mathcal{I}_n^{reg}[K]$ over $\mathcal{Z}_d^{\geq 0}(A)$ is unique and equals K_n^*.

Proof of Proposition 2. For i), this follows simply by using the fact that if $K(l,l') = \langle \eta(l), \eta(l')\rangle$ for some functions $\eta = (\eta_1, \ldots, \eta_d)$, then we have that

$$T_K[f](x) = \sum_{i=1}^d \eta_i(x)\int_0^1 \eta_i(y)f(y)\tilde{g}_n(y)\,dy = \sum_{i=1}^d (\eta_i)\otimes(\eta_i)^*$$

and consequently as $\mathrm{Tr}[\nu \otimes \nu^*] = \nu^*(\nu)$ and the trace is linear, we have that

$$\mathrm{Tr}[T_K] = \sum_{i=1}^d \int_{[0,1]} \eta_i(y)\eta_i(y)\tilde{g}_n(y)\,dy = \mathcal{I}_n^{\mathrm{reg}}[K].$$

Part ii) follows as $Z_{fr}^{\geq 0}(A)$ is free of A as a result of Lemma 52 [21].

For iii), as $\mathcal{I}_n^{\mathrm{reg}}[K]$ is simply the trace of the operator T_K, this will continuously extend to giving the trace on $\mathcal{Z}^{\geq 0}$, and more generally the positive trace-class operators on $L^2(\mu_n)$. This function is weak* lower semi-continuous as explained above.

To handle part iv), we note that if $\mathcal{I}_n^{\mathrm{reg}}[K] < \infty$, then T_K is trace-class, and consequently the operator T_K is also Hilbert-Schmidt, implying that $K \in L^2(\mu_n^{\otimes 2})$. We note that we have

$$\begin{gathered}0 < M^{-1} \leq \tilde{f}_n(l,l',x) \leq M < \infty, \qquad 0 < M^{-1} \leq \tilde{g}_n(l) \leq M < \infty, \\ |\ell(y,x) - \ell(y',x)| \leq L_\ell \max\{|y|,|y'|\}^{q-1}|y-y'|\end{gathered} \tag{50}$$

for all $l, l' \in [0,1]$, $y, y' \in \mathbb{R}$ and $x \in \{0,1\}$ (where $q = 1$ for the cross-entropy loss, and $q = 2$ for the squared loss), for some constants $M, L_\ell \in (0,\infty)$. It consequently therefore follows that for the cross-entropy loss we have that

$$\mathcal{I}_n[K] \leq 2M^3\int_{[0,1]^2}(\log(2) + |K(l,l')|)\,\tilde{g}_n(l)\tilde{g}_n(l')dldl' \ll \|K\|_{L^1(\mu_n^{\otimes 2})} \leq \|K\|_{L^2(\mu_n^{\otimes 2})} < \infty.$$

A similar argument holds for the squared loss function, after noting that $\ell(y,0) = y^2$ and $\ell(y,1) \leq 2(2+y^2)$ for all $y \in \mathbb{R}$. For the strict convexity, we note that this follows by the strict convexity of the loss functions $\ell(y,x)$, the positivity of the $\tilde{f}_n(l,l',x)$, and the fact that multiplying the $\ell(y,x)$ by $\tilde{f}_n(l,l',1)$ and $\tilde{f}_n(l,l',0)$, integrating, and then adding the two inequalities, will preserve the strict convexity.

By using the properties stated above in (50) we also can argue continuity of $\mathcal{I}_n[K]$, in that (recalling that $q = 1$ handles the cross-entropy loss, and $q = 2$ handles the squared loss)

$$\begin{aligned}
\big|\mathcal{I}_n[K] - \mathcal{I}_n[L]\big| &\leq ML_\ell \int_{[0,1]^2} \max\{|K(l,l')|, |L(l,l')|\}^{p-1} |K(l,l') - L(l,l')| \, dl dl' \\
&\leq 2M^3 \int_{[0,1]^2} \big(|K(l,l')| + |L(l,l')|\big)^{q-1} \big|K(l,l') - L(l,l')\big| \, \tilde{g}_n(l)\tilde{g}_n(l') \, dl dl' \\
&\leq 2M^3 \Big(\|K\|_{L^q(\mu_n^{\otimes 2})} + \|L\|_{L^q(\mu_n^{\otimes 2})}\Big)^{q-1} \|K - L\|_{L^q(\mu_n^{\otimes 2})} \\
&\leq 2M^3 \Big(\|K\|_{L^2(\mu_n^{\otimes 2})} + \|L\|_{L^2(\mu_n^{\otimes 2})}\Big)^{q-1} \|K - L\|_{L^2(\mu_n^{\otimes 2})} \\
&= 2M^3 \Big(\|T_K\|_{\text{HS}} + \|T_L\|_{\text{HS}}\Big)^{q-1} \|T_K - T_L\|_{\text{HS}} \\
&\leq 2M^3 \Big(\|T_K\|_1 + \|T_L\|_1\Big)^{q-1} \|T_K - T_L\|_1,
\end{aligned}$$

which also gives us part v); this is obtained by using (50) in the first line, the second by using the fact that $\tilde{g}_n$ is bounded below and that $\max\{|a|, |b|\} \leq |a| + |b|$; the third line by Hölder's inequality and the triangle inequality; the fourth line by Jensen's inequality; the fifth line by the identification between the L^2 norms of kernels and the Hilbert-Schmidt norm of their associated operators, and the last line by the fact that the trace norm upper bounds the Hilbert-Schmidt norm. In particular, $\mathcal{I}_n[K]$ is norm-continuous with respect to the norm of $L^2(\mu_n^{\otimes 2})$. This plus convexity implies that $\mathcal{I}_n[K]$ is weakly lower semi-continuous, in the sense of the weak topology on $L^2(\mu_n^{\otimes 2})$. The restriction of this topology to the trace-class operators is coarser than the weak* topology (by the definition of the weak topology), and therefore $\mathcal{I}_n[K]$ is also weak* lower semi-continuous, concluding the arguments for part iv).

For vi), this follows by using the above parts, the fact that the trace is linear over positive trace-class operators, and that the sum of convex and lower semi-continuous functions remain convex and lower semi-continuous respectively.

For vii), we first need to discuss some of the properties of the sets $\mathcal{Z}_d^{\geq 0}(A)$, $\mathcal{Z}_{\text{fr}}^{\geq 0}(A)$ and $\mathcal{Z}^{\geq 0}(A)$. We note that by the same argument in Proposition 47 of [21] that $\mathcal{Z}_d^{\geq 0}(A)$ is weak* closed, and that because of the facts a) $t\mathcal{Z}_d^{\geq 0}(A) \subset \mathcal{Z}_d^{\geq 0}(A)$ and b) $\mathcal{Z}_r^{\geq 0}(A) + \mathcal{Z}_s^{\geq 0}(A) = \mathcal{Z}_{r+s}^{\geq 0}(A)$, we can conclude that $\mathcal{Z}_{\text{fr}}^{\geq 0}(A) = \mathcal{Z}_{\text{fr}}^{\geq 0}$ - recall part ii) - is convex. As closures of convex sets are convex, it consequently follows that $\mathcal{Z}^{\geq 0}$ is convex and weak* closed. Noting that each of these sets contain 0, any minimizer K must satisfy

$$\xi_n \text{Tr}[T_K] \leq \mathcal{I}_n[K] + \xi_n \mathcal{I}_n^{\text{reg}}[K] \leq \mathcal{I}_n[0] + \xi_n \mathcal{I}_n^{\text{reg}}[0] = \mathcal{I}_n[0] \implies \text{Tr}[T_K] \leq \xi_n^{-1} \mathcal{I}_n[0].$$

As the set $\mathcal{B} := \{K : \text{Tr}[T_K] \leq \xi_n^{-1}\mathcal{I}_n[0]\}$ is weak* compact, it therefore follows that when minimizing over $\mathcal{Z}_d^{\geq 0}(A)$ and $\mathcal{Z}^{\geq 0}$, it suffices to minimize over the weak* compact sets $\mathcal{Z}_d^{\geq 0}(A) \cap \mathcal{B}$ and $\mathcal{Z}^{\geq 0} \cap \mathcal{B}$ respectively, and so by Weierstrass' theorem a minimizer must exist. As $\mathcal{I}_n[K] + \xi_n \mathcal{I}_n^{\text{reg}}[K]$ is strictly convex and $\mathcal{Z}^{\geq 0}$ is convex, we therefore also know that the minimizer over this set is unique.

To end with part viii), we highlight that in Appendix C.5, it is shown that when $\tilde{f}_n(l, l', 1)$, $\tilde{f}_n(l, l', 0)$ and $\tilde{g}_n(l)$ are piecewise constant, one can relate the minimization problem of minimizing $\mathcal{I}_n[K] + \xi_n \mathcal{I}_n^{\text{reg}}[K]$ over $\mathcal{Z}_d^{\geq 0}(A)$ to that of minimizing the function

$$\sum_{l,l' \in [\kappa]} p_n(l) p_n(l') \sum_x c_{f,n}(l, l', x) \ell(\langle \tilde{\omega}_l, \tilde{\omega}_{l'} \rangle, x) + \sum_{l \in [\kappa]} p_n(l) c_{g,n}(l) \|\tilde{\omega}_l\|_2^2$$

over $\tilde{\omega}_l$ for $l \in [\kappa]$ with $\|\tilde{\omega}_l\|_\infty \leq A$ for all A (see Appendix C.5 for a reminder of the relevant notation). In particular, in the case where we allow $d = \kappa$, and we relax the constraint on the $\tilde{\omega}_l$, if we write $\tilde{K}_{ll'} = \langle \tilde{\omega}_l, \tilde{\omega}_{l'} \rangle$, then we can write the above function as

$$\sum_{l,l' \in [\kappa(n)]} p_n(l) p_n(l') \sum_x c_{f,n}(l, l', x) \ell(\tilde{K}_{ll'}, x) + \sum_{l \in [\kappa(n)]} p_n(l) c_{g,n}(l) \tilde{K}_{ll},$$

which is a strictly convex function in the matrix $K_{ll'}$, and consequently has a unique minimizer over the cone of positive semi-definite matrices; call this matrix $\tilde{K}_n^*$. Supposing that $\tilde{K}_n^*$ is of rank $r(n) \leq \kappa$ (as the matrix is $\kappa \times \kappa$ dimensional and the rank is trivially less than the matrix dimension), if we write $\tilde{K}_n^* = \sum_{i=1}^{r}(n)\mu_i\phi_i\phi_i^T$ for some eigenvalues $\mu_i > 0$ and orthonormal eigenvectors $\phi_i \in \mathbb{R}^\kappa$, then we can identify K_n^* with $\tilde{K}_n^*$ via letting $K_n^* = \sum_{i=1}^{r(n)} \mu_i\psi_i(l)\psi_i(l')$ where $\psi_i(l) = \phi_{ij}$ for $l \in A_j$. We now highlight that one trivially has that $\|\phi_i\|_\infty \leq 1$ for all i, and moreover that as every row and column sum (ignoring the diagonal) is bounded above by $(\kappa-1)\|\tilde{K}_n^*\|_\infty$, by the Gershgorin circle theorem the eigenvalues are bounded above by $(\kappa-1)\|\tilde{K}_n^*\|_\infty$ also. Consequently, as soon as $d \geq r(n)$ and $A \geq (\kappa-1)\|\tilde{K}_n^*\|_\infty$, $K_n^* \in \mathcal{Z}_d^{\geq 0}(A)$, and as a result we have that the minima of $\mathcal{I}_n[K] + \xi_n\mathcal{I}_n^{\text{reg}}[K]$ over $\mathcal{Z}_d^{\geq 0}(A)$ is unique and equals K_n^*. □

As the above theorem shows that $\mathcal{I}_n[K] + \xi_n\mathcal{I}_n^{\text{reg}}[K]$ is a strictly convex function, well defined for all symmetric kernels K corresponding to positive, self-adjoint, trace class operators $L^2(\mu_n) \to L^2(\mu_n)$ via the identification $K \to T_K$ given in (48), we briefly discuss here the corresponding KKT conditions for constrained minimization.

Proposition 3. *Let $\mathcal{C}$ be a weak* closed set of positive, symmetric, trace class kernels. Then L is the unique minima of $\mathcal{I}_n[K] + \xi_n\mathcal{I}_n^{reg}[K]$ over $\mathcal{C}$ if and only if there exists some $V \in \mathcal{B}(L^2(\mu_n))$ such that*

$$\text{Tr}(VT_L) = \mathcal{I}_n^{reg}[K], \qquad \|V\|_{op} \leq 1, \qquad \text{Tr}\big((T_\nabla + \xi_n V)(T_K - T_L)\big) \geq 0 \text{ for all } K \in \mathcal{C},$$

where we identify symmetric kernels $K \in L^2(\mu_n^{\otimes 2})$ with operators $L^2(\mu_n) \to L^2(\mu_n)$ as in (48)*, and write T_∇ for the bounded operator $L^2(\mu_n) \to L^2(\mu_n)$ with kernel*

$$\nabla\mathcal{I}_n[K] = \sum_{x\in\{0,1\}} \frac{\tilde{f}_n(l,l',x)\ell'(K(l,l'),x)}{\tilde{g}_n(l)\tilde{g}_n(l')},$$

where $\ell'(y,x)$ is the derivative of $\ell(y,x)$ with respect to y.

Proof of Proposition 3. We begin by deriving the subgradient for both $\mathcal{I}_n[K]$ and $\mathcal{I}_n^{\text{reg}}[K]$, and then use the rules of subgradient calculus to obtain the KKT conditions. For $\mathcal{I}_n[K]$, note that we can write

$$\mathcal{I}_n[K] := \int_{[0,1]^2} \sum_{x\in\{0,1\}} \frac{\tilde{f}_n(l,l',x)\ell(K(l,l'),x)}{\tilde{g}_n(l)\tilde{g}_n(l')}\, \tilde{g}_n(l)\tilde{g}_n(l')\, dldl' \tag{51}$$

and so the subgradient (in terms of the operator) is a singleton, say T_∇, whose sole element is the operator with kernel given by the Fréchet derivative of $\mathcal{I}_n[K]$

$$\nabla\mathcal{I}_n[K](l,l') = \sum_{x\in\{0,1\}} \frac{\tilde{f}_n(l,l',x)\ell'(K(l,l'),x)}{\tilde{g}_n(l)\tilde{g}_n(l')} \tag{52}$$

[e.g 5, Proposition 2.53]. As for $\mathcal{I}_n^{\text{reg}}[K]$, we recall that this equals $\text{Tr}[T_K]$, i.e the trace norm of T_K, as K is positive. Because the dual space of $S(L^2(\mu_n))$ is the space of bounded operators $L^2(\mu_n) \to L^2(\mu_n)$ equipped with norm $\|\cdot\|_{\text{op}}$, we have that

$$\partial\mathcal{I}_n^{\text{reg}}[K] = \big\{V \in \mathcal{B}(L^2(\mu_n)) \,:\, \text{Tr}(VT_K) = \mathcal{I}_n^{\text{reg}}[K], \|V\|_{\text{op}} \leq 1\big\} \tag{53}$$

[e.g 2, Theorem 7.57]. Combining the two subgradients together says that L is an optimizer to $\mathcal{I}_n[K] + \xi_n\mathcal{I}_n^{\text{reg}}[K]$ over $\mathcal{C}$ if and only if there exists some $V \in \mathcal{B}(L^2(\mu_n))$ such that

$$\text{Tr}(VT_L) = \mathcal{I}_n^{\text{reg}}[K], \qquad \|V\|_{\text{op}} \leq 1, \qquad \text{Tr}\big((T_\nabla + \xi_n V)(T_K - T_L)\big) \geq 0 \text{ for all } K \in \mathcal{C} \tag{54}$$

as stated. □

With this, we now state the full version of Theorem 2, complete with regularity conditions.

Theorem 7. *Suppose that Assumptions 2 and 3 hold and that $\xi_n = O(1)$. Write $\mathcal{Z}^{\geq 0} = \text{cl}\big(\cup_{d\geq 1} \mathcal{Z}_d^{\geq 0}(A)\big)$ for the closure of the union of the $\mathcal{Z}_d^{\geq 0}(A)$ with respect to the weak* topology on the*

trace-class operators $L^2(\mu_n) \to L^2(\mu_n)$ as described in Proposition 2. For each n, let K_n^* denote the unique minimizer to the optimization problem

$$\min_{K \in \mathcal{Z}^{\geq 0}(A)} \big\{ \mathcal{I}_n[K] + \xi_n \mathcal{I}_n^{reg}[K] \big\},$$

and assume that the K_n^* are uniformly bounded in $L^\infty([0,1]^2)$. Moreover, suppose that either

(I) on the same partition $\mathcal{Q}$ as given in Assumption 3a), we have that K_n^* is piecewise constant on $\mathcal{Q} \times \mathcal{Q}$;

(II) the K_n^* are all Hölder$([0,1], \beta^*, L^*)$ for some constants β^* and L^*.

Then there exists A' (see Lemma 7 and Lemma 8) such that whenever $A_1, A_2 \geq A'$, for any sequence of minimizers

$$\widehat{\boldsymbol{\omega}}_n \in \operatorname*{arg\,min}_{\boldsymbol{\omega}_n \in ([-A_1, A_1]^d)^n} \big\{ \mathcal{R}_n(\boldsymbol{\omega}_n) + \xi_n \mathcal{R}_n^{reg}(\boldsymbol{\omega}_n) \big\} \text{ such that } \max_{i,j} \big| \langle \widehat{\omega}_i, \widehat{\omega}_j \rangle \big| \leq A_2$$

we have that under condition (II) that

$$\frac{1}{n^2} \sum_{i,j \in [n]} \Big(\langle \widehat{\omega}_i, \widehat{\omega}_j \rangle - K_n^*(\lambda_i, \lambda_j) \Big)^2 = O_p\Big(r_n + d^{-\beta *} + \Big(\frac{\log(n)}{n} \Big)^{\min\{\beta, \beta^*\}/2} \Big),$$

where r_n is the relevant rate of convergence in Theorem 1. In particular, there exists a sequence of embedding dimensions $d = d(n)$ such that the above bound is $o_p(1)$. Under condition (I), the above rate of convergence can be improved as follows: there exists some constant $r \leq \kappa$ such that, as soon as $d \geq r$, we have that the above bound is of the order $O_p(r_n)$ only. In particular, as soon as $d \geq r$, the above bound is $o_p(1)$.

Remark 4. *The conditions on K_n^* are given in order to give explicit rates of convergence; in order to only argue that we obtain consistency of the bound given above, it suffices to have that the K_n^* are equicontinuous for each n. Moreover, this is only necessary in order to relate the minimal values of the $\langle \widehat{\omega}_i, \widehat{\omega}_j \rangle$ directly to the values of $K_n^*(\lambda_i, \lambda_j)$; we can still obtain weaker notions of consistency (see e.g.* (D)*) if we do not impose any continuity requirements. With regards to the assumption that the infinity norm of the matrix $\langle \widehat{\omega}_i, \widehat{\omega}_j \rangle$ is bounded with n, this could be imposed as a constraint in Theorem 1 to guarantee such a pair of minimizers; as highlighted in Remark 3, this can lead to improved dependence on the dimension d. We highlight that as under the given assumptions on the $\tilde{f}_n(l, l', 1)$ and $\tilde{f}_n(l, l', 0)$, the unconstrained minimizer when $\xi_n = 0$ is uniformly bounded in $L^\infty([0,1]^2)$, and so we do not consider these assumptions (both on K_n^* and the gram matrix of the embedding vectors) to be restrictive.*

Remark 5. *We highlight that we usually expect $\beta = \beta^*$; for example, see Theorem 5 for an example with the squared loss.*

Remark 6. *We briefly discuss the rates of convergence of the above estimator when in the dense regime and using the squared loss, as in this setting the bound we obtain naturally corresponds to the guarantees given in the graphon estimation literature. In particular, when $\tilde{f}_n(l, l', 1)$, $\tilde{f}_n(l, l', 0)$ and $\tilde{g}_n(l)$ are constant (i.e, free of l), Theorem 5 guarantees us that the minima of $\mathcal{I}_n[K] + \xi_n \mathcal{I}_n^{reg}[K]$ corresponds to a version of the original generating graphon W whose singular values have been subject to a soft-thresholding operator, and we can take $\beta^* = \beta$ also.*

In such a scenario, we then note that if we also take Remark 3 into account, then the rate of convergence equals

$$s_n + \Big(\frac{d}{n} \Big)^{1/2} + \Big(\frac{\log n}{n^{2\beta/(1+2\beta)}} \Big)^{1/2} + d^{-\beta} + \Big(\frac{\log n}{n} \Big)^{\beta/2}.$$

By choosing the embedding dimension d optimally so that $d = O(n^{1/(1+2\beta)})$, and noting that the $\big(\log n / n^{2\beta/(1+2\beta)} \big)^{1/2}$ term is of a slower order than the $\big(\log n / n \big)^{\beta/2}$ term, we end up with a rate of convergence

$$s_n + \Big(\frac{\log n}{n^{2\beta/(1+2\beta)}} \Big)^{1/2}.$$

Up to logarithmic factors and the sampling term, this is a square root of the rate of convergence of the UVST procedure [66], which is itself a square root of the minimax rates of estimation [22]. We

suspect that the difference with the rates achieved in [66] occurs due to our approach of looking at the rates of convergence between the empirical and population risks, rather than being able to work directly with the original objective at all times. It would be interesting to see whether the rates of convergence can be improved so that, up to the sampling term, we end up with the same rates of convergence as in [66].

Proof of Theorem 7. The idea of the proof is to associate a kernel $\widehat{K}$ to a minimizer $\widehat{\boldsymbol{\omega}}_n$ of $\mathcal{R}_n(\boldsymbol{\omega}_n) + \xi_n \mathcal{R}_n^{\text{reg}}(\boldsymbol{\omega}_n)$ over $([-A, A]^d)^n$, and then argue from the uniform convergence results developed in the proof of Theorem 1 that this requires $\widehat{K}$ to be close to the minimizer of $\mathcal{I}_n[K] + \xi_n \mathcal{I}_n^{\text{reg}}[K]$. Consequently, we can then use the curvature of this function about its minima to derive consistency guarantees.

To associate a kernel K to a collection of embedding vectors $\boldsymbol{\omega}_n$, we begin by writing $\lambda_{n,(i)}$ for the associated order statistics of $\boldsymbol{\lambda}_n = (\lambda_1, \dots, \lambda_n)$, and let π_n be the mapping which sends i to the rank of λ_i. We then define the sets

$$A_{n,i} = \Big[\frac{i - 1/2}{n+1}, \frac{i+1/2}{n+1}\Big] \qquad \text{for } i \in [n],$$

and define the sequence of functions

$$\widehat{K}_n(l, l') = \langle \widehat{\eta}(l), \widehat{\eta}(l') \rangle \qquad \text{where} \qquad \widehat{\eta}(l) = \begin{cases} \widehat{\omega}_i & \text{if } l \in A_{n,\pi_n(i)}, \\ 0 & \text{otherwise.} \end{cases}$$

for any sequence $\widehat{\boldsymbol{\omega}}_n$ of minimizers to $\mathcal{R}_n(\boldsymbol{\omega}_n) + \xi_n \mathcal{R}_n^{\text{reg}}(\boldsymbol{\omega}_n)$. The idea of the proof is to then focus on upper and lower bounding the quantity

$$\big\{\mathcal{I}_n[\widehat{K}_n] + \xi_n \mathcal{I}_n^{\text{reg}}[\widehat{K}_n]\big\} - \big\{\mathcal{I}_n[K_n^*] + \xi_n \mathcal{I}_n^{\text{reg}}[K_n^*]\big\},$$

where K_n^* is the minimizer of $\mathcal{I}_n[K] + \xi_n \mathcal{I}_n^{\text{reg}}[K]$ over $\mathcal{Z}^{\geq 0}(A)$.

Step 1: Bounding from above. Begin by noting from the triangle inequality we have that

$$\begin{aligned}
&\big\{\mathcal{I}_n[\widehat{K}_n] + \xi_n \mathcal{I}_n^{\text{reg}}[\widehat{K}_n]\big\} - \big\{\mathcal{I}_n[K_n^*] + \xi_n \mathcal{I}_n^{\text{reg}}[K_n^*]\big\} \\
&\qquad \leq \Big|\big\{\mathcal{I}_n[\widehat{K}_n] + \xi_n \mathcal{I}_n^{\text{reg}}[\widehat{K}_n]\big\} - \big\{\mathcal{R}_n(\widehat{\boldsymbol{\omega}}_n) + \xi_n \mathcal{R}_n^{\text{reg}}(\widehat{\boldsymbol{\omega}}_n)\big\}\Big| \qquad \text{(I)} \\
&\qquad + \Big|\big\{\mathcal{R}_n[\widehat{\boldsymbol{\omega}}_n] + \xi_n \mathcal{R}_n^{\text{reg}}(\widehat{\boldsymbol{\omega}}_n)\big\} - \min_{K \in \mathcal{Z}_d^{\geq 0}(A)} \big\{\mathcal{I}_n[K] + \xi_n \mathcal{I}_n^{\text{reg}}[K]\big\}\Big| \qquad \text{(II)} \\
&\qquad + \Big|\min_{K \in \mathcal{Z}_d^{\geq 0}(A)} \big\{\mathcal{I}_n[K] + \xi_n \mathcal{I}_n^{\text{reg}}[K]\big\} - \min_{K \in \mathcal{Z}^{\geq 0}(A)} \big\{\mathcal{I}_n[K] + \xi_n \mathcal{I}_n^{\text{reg}}[K]\big\}\Big|. \qquad \text{(III)}
\end{aligned}$$

We want to bound each of the terms (I), (II) and (III). By using Lemma 6, Theorem 1 and Lemma 7 respectively, we end up being able to bound the above quantity by $O_p(q_n)$, where

$$q_n = \begin{cases} r_n & \text{if (I) holds} \\ r_n + \big(\log(n)/n\big)^{\beta/2} + d^{-\beta^*} & \text{if (II) holds.} \end{cases}$$

Step 2: Bounding from below. Let q_n denote the upper bound on the rate of convergence of $\big\{\mathcal{I}_n[\widehat{K}_n] + \xi_n \mathcal{I}_n^{\text{reg}}[\widehat{K}_n]\big\} - \big\{\mathcal{I}_n[K_n^*] + \xi_n \mathcal{I}_n^{\text{reg}}[K_n^*]\big\}$ as developed above. Then by Lemma 8, we have that

$$\int_{[0,1]^2} \Big(\widehat{K}_n(l, l') - K_n^*(l, l')\Big)^2 \, dl dl' = O_p(q_n).$$

If we then define the function

$$\overline{K}_n^*(l, l') = \begin{cases} K_n^*(\lambda_i, \lambda_j) & \text{if } (l, l') \in A_{n,\pi_n(i)} \times A_{n,\pi_n(j)}, \\ 0 & \text{otherwise} \end{cases}$$

then by the same arguments as in the proof of Lemma 6 we get that

$$\int_{[0,1]^2} \big(\overline{K}_n^*(l, l') - K_n^*(l, l')\big)^2 \, dl dl' = \begin{cases} O_p(n^{-1/2}) & \text{if (I) holds} \\ O_p\big((\log(n)/n)^{\beta^*/2}\big) & \text{if (II) holds.} \end{cases}$$

Consequently, as a result of the triangle inequality we get that

$$\frac{1}{(n+1)^2}\sum_{i,j\in[n]}\big(K_n^*(\lambda_i,\lambda_j)-\langle\widehat{\omega}_i,\widehat{\omega}_j\rangle\big)^2$$
$$=\int_{[0,1]^2}\big(\overline{K}_n^*(l,l')-\widehat{K}_n(l,l')\big)^2\,dldl'=\begin{cases}O_p(q_n) & \text{if (I) holds}\\ O_p\big(q_n+(\log(n)/n)^{\beta^*/2}\big) & \text{if (II) holds,}\end{cases}$$

giving the desired result. □

D.1 Additional lemmata

Lemma 6. *Under the assumptions and notation of Theorem 7, we have that*

$$\Big|\big\{\mathcal{I}_n[\widehat{K}_n]+\xi_n\mathcal{I}_n^{reg}[\widehat{K}_n]\big\}-\big\{\mathcal{R}_n(\widehat{\boldsymbol{\omega}}_n)+\xi_n\mathcal{R}_n^{reg}(\widehat{\boldsymbol{\omega}}_n)\big\}\Big|=O_p\big(r_n+(\log(n)/n)^{\beta/2}\big),$$

where r_n is the convergence rate in Theorem 1 when condition (II) holds. When condition (I) holds, the rate of convergence can be improved to be simply $O_p(r_n)$.

Proof of Lemma 6. We begin by handling what occurs when condition (II) of Theorem 7 holds, and then detail what changes when condition (I) holds instead.

Begin by defining the quantities

$$\tilde{c}_{f,n}(i,j,x):=\frac{1}{|A_{n,\pi_n(i)}||A_{n,\pi_n(j)}|}\int_{A_{n,\pi_n(i)}\times A_{n,\pi_n(j)}}\tilde{f}_n(l,l',x)\,dldl',$$
$$\tilde{c}_{g,n}(l)=\frac{1}{|A_{n,\pi_n(i)}|}\int_{A_{n,\pi_n(i)}}\tilde{g}_n(l)\,dl,$$

and note that as

$$\max_{i\in[n]}\Big|\lambda_{n,(i)}-\frac{i}{n+1}\Big|=O_p\Big(\Big(\frac{\log(2n)}{n}\Big)^{1/2}\Big)$$

[by e.g 45, Theorem 2.1], we get that

$$\begin{aligned}&\big|\tilde{c}_{f,n}(i,j,x)-\tilde{f}_n(\lambda_i,\lambda_j,x)\big|\\&\quad\le\frac{1}{|A_{n,\pi_n(i)}||A_{n,\pi_n(j)}|}\int_{A_{n,\pi_n(i)}\times A_{n,\pi_n(j)}}\big|\tilde{f}_n(l,l',x)-\tilde{f}_n(\lambda_{n,(\pi_n(i))},\lambda_{n,(\pi_n(j))},x)\big|\,dldl'\\&\quad\le L\sup_{(l,l')\in A_{n,\pi_n(i)}\times A_{n,\pi_n(j)}}\|(l,l')-(\lambda_{n,(\pi_n(i))},\lambda_{n,(\pi_n(j))})\|_2^{\beta}\\&\quad\le L2^{\beta/2}\Big(\frac{1}{2n}+\max_{i\in[n]}\big|\lambda_{n,(i)}-\frac{i}{n+1}\big|\Big)^{\beta}=O_p\Big(\Big(\frac{\log n}{n}\Big)^{\beta/2}\Big),\end{aligned}$$

uniformly for all i,j, and similarly

$$\big|\tilde{c}_n(i)-\tilde{g}_n(\lambda_i)\big|=O_p\Big(\Big(\frac{\log n}{n}\Big)^{\beta/2}\Big)$$

uniformly over i. Using the fact that $\widehat{K}_n$ is piecewise constant, we can then write

$$\begin{aligned}&\mathcal{I}_n[\widehat{K}_n]+\xi_n\mathcal{I}_n^{\text{reg}}[\widehat{K}_n]\\&\quad=\frac{1}{(n+1)^2}\sum_{(i,j)\in[n]}\sum_{x\in\{0,1\}}\ell(\langle\widehat{\omega}_i,\widehat{\omega}_j\rangle,x)\tilde{c}_{f,n}(i,j,x)+\xi_n\sum_{i\in[n]}\|\widehat{\omega}_i\|_2^2\tilde{c}_{g,n}(i)+\frac{2(n-1)c_\ell}{(n+1)^2},\end{aligned}$$

where c_ℓ is a constant which depends on the choice of the loss function. Introducing the function (compare with $\mathbb{E}[\widehat{\mathcal{R}}_{n,(1)}^{\mathcal{P}_n}(\boldsymbol{\omega}_n)\,|\,\boldsymbol{\lambda}_n]$ from (40))

$$\mathbb{E}[\widehat{\mathcal{R}}_{n,(1)}(\boldsymbol{\omega}_n)\,|\,\boldsymbol{\lambda}_n]:=\frac{1}{n^2}\sum_{i,j\in[n]}\sum_{x\in\{0,1\}}\tilde{f}_n(\lambda_i,\lambda_j,x)\ell(\langle\omega_i,\omega_j\rangle,x),$$

it follows that

$$\begin{aligned}
&\big|\big\{\mathcal{I}_n[\widehat{K}_n] + \xi_n \mathcal{I}_n^{\text{reg}}[\widehat{K}_n]\big\} - \big\{\mathcal{R}_n(\widehat{\boldsymbol{\omega}}_n) + \xi_n \mathcal{R}_n^{\text{reg}}(\widehat{\boldsymbol{\omega}}_n)\big\}\big| \\
&\qquad \le \Big|\frac{1}{(n+1)^2}\sum_{i,j\in[n]}\sum_{x\in\{0,1\}} \ell(\langle\widehat{\omega}_i,\widehat{\omega}_j\rangle, x)\big\{\tilde{c}_{f,n}(i,j,x) - \tilde{f}_n(\lambda_i,\lambda_j,x)\big\} \\
&\qquad\qquad + \frac{\xi_n}{n+1}\sum_{i\in[n]} \|\widehat{\omega}_i\|_2^2\big\{\tilde{c}_{g,n}(l) - \tilde{g}_n(\lambda_i)\big\}\Big| \qquad \text{(A)} \\
&\qquad + \big|\big\{\mathbb{E}[\widehat{\mathcal{R}}_{n,(1)}(\widehat{\boldsymbol{\omega}}_n)\,|\,\boldsymbol{\lambda}_n] + \xi_n \widehat{\mathcal{R}}_n^{\text{reg}}(\widehat{\boldsymbol{\omega}}_n)\big\} - \big\{\mathcal{R}_n(\widehat{\boldsymbol{\omega}}_n) + \xi_n \mathcal{R}_n^{\text{reg}}(\widehat{\boldsymbol{\omega}}_n)\big\}\big| \qquad \text{(B)} \\
&\qquad + O(n^{-1})\big\{\mathbb{E}[\widehat{\mathcal{R}}_{n,(1)}(\widehat{\boldsymbol{\omega}}_n)\,|\,\boldsymbol{\lambda}_n] + \xi_n \widehat{\mathcal{R}}_n^{\text{reg}}(\widehat{\boldsymbol{\omega}}_n)\big\} + O(n^{-1}). \qquad \text{(C)}
\end{aligned}$$

From the proof[2] of Theorem 1, we know that the (B) term is of the order $O_p(r_n)$, and consequently via the uniform convergence bounds developed throughout the proof, this will also imply that $\big\{\mathbb{E}[\widehat{\mathcal{R}}_{n,(1)}(\widehat{\boldsymbol{\omega}}_n)\,|\,\boldsymbol{\lambda}_n] + \xi_n \widehat{\mathcal{R}}_n^{\text{reg}}(\widehat{\boldsymbol{\omega}}_n)\big\} = O_p(1)$ and consequently the term in (C) will be of the order $O_p(n^{-1})$. For (A), we begin by noting that (A) can be bounded via the triangle inequality and the observations above by

$$\text{(A)} \le \Big(\frac{1}{n^2}\sum_{i,j\in[n]}\sum_{x\in\{0,1\}} \ell(\langle\widehat{\omega}_i,\widehat{\omega}_j\rangle, x) + \frac{1}{n}\sum_{i\in[n]}\|\widehat{\omega}_i\|_2^2\Big)\cdot O_p\Big(\Big(\frac{\log n}{n}\Big)^{\beta/2}\Big).$$

To conclude, we just need to argue that

$$\frac{1}{n^2}\sum_{i,j\in[n]}\sum_{x\in\{0,1\}} \ell(\langle\widehat{\omega}_i,\widehat{\omega}_j\rangle, x) + \frac{1}{n}\sum_{i\in[n]}\|\widehat{\omega}_i\|_2^2 = O_p(1).$$

To see this, we note that this simply follows by using the fact that $\big\{\mathbb{E}[\widehat{\mathcal{R}}_{n,(1)}(\widehat{\boldsymbol{\omega}}_n)\,|\,\boldsymbol{\lambda}_n] + \xi_n \widehat{\mathcal{R}}_n^{\text{reg}}(\widehat{\boldsymbol{\omega}}_n)\big\} = O_p(1)$ (as argued above) and the fact that the $\tilde{f}_n(l,l',1)$, $\tilde{f}_n(l,l',0)$ and $\tilde{g}_n$ are assumed to be uniformly bounded below by M^{-1}.

When condition (I) holds instead, we need to change the style of argument. Note that when $\mathcal{Q} = (A_1,\ldots,A_\kappa)$, if we define the sets

$$N_{\lambda,n,k} := \{j \,:\, \lambda_j \in A_k\}, \quad N_{A,n,k} = \{j \,:\, A_{n,\pi_n(j)}\ A_k\}$$

$$M_{n,k} = N_{\lambda,n,k} \cap N_{A,n,k}, \quad M_n = \bigcup_{k=1}^{\kappa} M_{n,k},$$

then by Theorem 63 of [21], we have that $|M_n| \ge n - O_p(\sqrt{n})$, $|M_n^c| \le O_p(\sqrt{n})$. To make use of this, note that

$$\big|\tilde{c}_{f,n}(i,j,x) - \tilde{f}_n(\lambda_i,\lambda_j,x)\big| = \begin{cases} 0 & \text{if } i,j \in M_n \\ M & \text{otherwise,} \end{cases}$$

and also that

$$\big|\tilde{c}_n(i) - \tilde{g}_n(\lambda_i)\big| = \begin{cases} 0 & \text{if } i \in M_n \\ M & \text{otherwise.} \end{cases}$$

Writing $c_{\ell,2} = \max\{\ell(A_2,1), \ell(A_2,0), \ell(-A_2,0), \ell(-A_2,1)\}$, the bound in (A) is replaced by

$$\text{(A)} \le M\Big(c_{\ell,2}\frac{|M_n^c|^2 + 2|M_n||M_n^c|}{(n+1)^2} + \frac{\xi_n|M_n^c|}{n+1}\Big) = O_p(n^{-1/2}),$$

and so the argument progresses through as before, except that we can drop the $(\log n/n)^{\beta/2}$ term in the overall rate of convergence. □

Lemma 7. *Under the assumptions and notation of Theorem 7, there exists A' such that whenever $A \ge A'$, we have, under condition (II) of Theorem 7, that*

$$\sup_{n\ge 1}\Big|\min_{K\in\mathcal{Z}_d^{\ge 0}(A)}\big\{\mathcal{I}_n[K] + \xi_n\mathcal{I}_n^{reg}[K]\big\} - \min_{K\in\mathcal{Z}^{\ge 0}}\big\{\mathcal{I}_n[K] + \xi_n\mathcal{I}_n^{reg}[K]\big\}\Big| = O(d^{-\beta^*}).$$

[2]We note that the step where the 'diagonal term' of including/excluding the sums of $\ell(\langle\omega_i,\omega_j\rangle, x)$ can be carried out before or after the stepping approximation step.

When condition (I) holds instead, then there exists $r \leq \kappa$ and $A' < \infty$ such that, as soon as $d \geq r$ and $A \geq A'$, we have that

$$\sup_{n\geq 1}\Big|\min_{K\in\mathcal{Z}_d^{\geq 0}(A)}\big\{\mathcal{I}_n[K]+\xi_n\mathcal{I}_n^{reg}[K]\big\}-\min_{K\in\mathcal{Z}^{\geq 0}}\big\{\mathcal{I}_n[K]+\xi_n\mathcal{I}_n^{reg}[K]\big\}\Big|=0.$$

Proof of Lemma 7. We begin with the argument under condition (II) first, and then highlight how the details change when condition (I) holds instead. Note that by the spectral theorem for self-adjoint compact operators, we can write for each n the eigendecomposition

$$K_n^*(l,l')=\sum_{i=1}^{\infty}\lambda_i(K_n^*)\psi_{n,i}(l)\psi_{n,i}(l')$$

where the $\lambda_i(K_n^*)$ are non-negative, monotone decreasing in i for each n, and satisfy the bound $\lambda_i(K_n^*) = O(d^{-(1+\beta^*)})$ [52], and are uniformly bounded above by $\|K_n^*\|_{L^2(\mu_n^{\otimes 2})} \leq M^2\sup_{n\geq 1}\|K_n^*\|_\infty$. As for the eigenfunctions, we note that they are orthonormal in that $\langle\psi_{n,i},\psi_{n,j}\rangle_{L^2(\mu_n)}=\delta_{ij}$. Moreover, as the image of the operator $T_{K_n^*}$ under the unit $L^2(\mu_n)$ ball lies within the class of Hölder($[0,1]$, β^*, L^*) functions, the $\psi_{n,i}$ are each Hölder($[0,1]$, β^*, L^*), and as they are uniformly bounded in $L^2(\mu_n)$, they will also be uniformly bounded (across i and n) in $L^\infty([0,1])$ too (see e.g. Lemma 10). Consequently, writing

$$K_{n,d}^*(l,l')=\sum_{i=1}^{d}\lambda_i(K_n^*)\psi_{n,i}(l)\psi_{n,i}(l')$$

for the best rank d approximation to K_n^*, it follows that $K_{n,d}^*\in\mathcal{Z}_d^{\geq 0}(A)$ for any $A\geq A' = M\sqrt{\sup_{n\geq 1}\|K_n^*\|_\infty}\cdot\sup_{n,i}\|\psi_{n,i}\|_\infty$. As a result, we have that

$$\begin{aligned}\min_{K\in\mathcal{Z}_d^{\geq 0}(A)}\big\{\mathcal{I}_n[K]+\xi_n\mathcal{I}_n^{\text{reg}}[K]\big\}-\min_{K\in\mathcal{Z}^{\geq 0}}\big\{\mathcal{I}_n[K]+\xi_n\mathcal{I}_n^{\text{reg}}[K]\big\}&\\ \leq\big\{\mathcal{I}_n[K_{n,d}^*]+\xi_n\mathcal{I}_n^{\text{reg}}[K_{n,d}^*]\big\}-\big\{\mathcal{I}_n[K_n^*]+\xi_n\mathcal{I}_n^{\text{reg}}[K_n^*]\big\}.&\end{aligned}$$

In order to obtain the final bound, we then note that by the local-Lipschitz property derived in Proposition 2v), in addition to the fact that the trace is linear and equals the sum of the eigenvalues of the operator, we get that (where we use $\lesssim$ to hide unimportant constants)

$$\begin{aligned}&\big\{\mathcal{I}_n[K_{n,d}^*]+\xi_n\mathcal{I}_n^{\text{reg}}[K_{n,d}^*]\big\}-\big\{\mathcal{I}_n[K_n^*]+\xi_n\mathcal{I}_n^{\text{reg}}[K_n^*]\big\}\\ &\quad\lesssim(2\|K_n^*\|_{L^2(\mu_n^{\otimes 2})})^{q-1}\cdot\|K_{n,d}^*-K_n^*\|_{L^2(\mu_n^{\otimes 2})}+\xi_n\big|\text{Tr}[T_{K_n^*}-T_{K_{n,d}^*}]\big|\\ &\quad=O\Big(\sup_{n\geq 1}\|K_n^*\|_\infty^{q-1}\Big(\sum_{i=d+1}^{\infty}d^{-2(1+\beta^*)}\Big)^{1/2}+\sum_{i=d+1}^{\infty}d^{-(1+\beta^*)}\Big)=O(d^{-\beta^*}),\end{aligned}$$

as desired, noting that the bound on the RHS holds uniformly in n.

We highlight that in the case where condition (I) holds, we know by the last part of Proposition 2 that, for each n, there exists $r(n)\leq\kappa$ such that once $d\geq r(n)$ and $A\geq(\kappa-1)\|K_n^*\|_\infty$, the minima of $\mathcal{I}_n[K]+\xi_n\mathcal{I}_n^{\text{reg}}[K]$ over $\mathcal{Z}_d^{\geq 0}(A)$ equals the minima over the set $\mathcal{Z}^{\geq 0}$. Consequently, under the assumptions stated, letting $r=\tilde{r}=\sup_{n\geq 1}r(n)\leq\kappa$ and $A'=(\kappa-1)\sup_{n\geq 1}\|K_n^*\|_\infty$ gives the stated result. □

Lemma 8. *When $\ell(y,x)$ is the cross-entropy loss, under the assumptions and notation of Theorem 7, for any $K\in\mathcal{Z}^{\geq 0}$ such that $\|K\|_\infty<\infty$, we have that*

$$\big\{\mathcal{I}_n[K]+\xi_n\mathcal{I}_n^{reg}[K]\big\}-\big\{\mathcal{I}_n[K_n^*]+\xi_n\mathcal{I}_n^{reg}[K_n^*]\big\}\geq\frac{1}{2M}\int_{[0,1]^2}\big(K_n^*(l,l')-K(l,l')\big)^2\,dldl'$$

for some constant M which depends on $C_M:=\max\{\|K\|_\infty,\sup_n\|K_n^*\|_\infty\}<\infty$; in particular one can take $M=(e^{C_M}/(1+e^{C_M})^2)^{-1}$. When $\ell(y,x)$ is the squared loss, one can relax the requirement that $\|K\|_\infty<\infty$, and can take $M=1/2$ instead.

Proof of Lemma 8. We give the details for the cross-entropy loss, as the argument for the squared loss is the same, except for that the requirement that $\|K\|_\infty < \infty$ can be dropped. To begin, we note that for any y and y' such that $|y|, |y'| \leq A$ for some constant A, we have that $\ell''(y,x) = e^y/(1+e^y)^2 \leq e^A/(1+e^A)^2 > 0$, and consequently $\ell(y,x)$ is strongly convex in y on the domain $|y| \leq A$ for all $x \in \{0,1\}$. As a result, we have the inequality

$$\ell(y,x) \geq \ell(y',x) + (y-y')\ell'(y',x) + \frac{e^A}{2(1+e^A)^2}(y-y')^2$$

for $x \in \{0,1\}$ and all y, y' with $|y|, |y'| \leq A$. After multiplying the above inequality by the $\tilde{f}_n(l,l',1)$ and $\tilde{f}_n(l,l',0)$ separately, adding the two inequalities together, and integrating, we obtain the inequality

$$\mathcal{I}_n[K] \geq \mathcal{I}_n[K'] + \int_{[0,1]^2} \nabla \mathcal{I}_n[K'](K(l,l') - K'(l,l'))\, dl dl' + \frac{1}{2M}\int_{[0,1]^2} \Big(K(l,l') - K'(l,l')\Big)^2 dl dl'$$

for any $K, K' \in \mathcal{Z}^{\geq 0}(A)$ for which $\|K\|_\infty, \|K'\|_\infty < \infty$, where M depends on the value of $C_M := \max\{\|K\|_\infty, \|K'\|_\infty\}$; in particular, we have that $M = (e^{C_M}/(1+e^{C_M})^2)^{-1}$. Note that under our assumptions, the K_n^* are uniformly bounded in $L^\infty([0,1]^2)$, and consequently it follows that for any $K \in \mathcal{Z}^{\geq 0}(A)$ which is bounded in $L^\infty([0,1]^2)$ that

$$\begin{aligned}
&\big\{\mathcal{I}_n[K] + \xi_n \mathcal{I}_n^{\text{reg}}[K]\big\} - \big\{\mathcal{I}_n[K_n^*] + \xi_n \mathcal{I}_n^{\text{reg}}[K_n^*]\big\} \\
&\overset{(a)}{\geq} \int_{[0,1]^2} \nabla \mathcal{I}_n[K_n^*](K(l,l') - K_n^*(l,l'))\, dl dl' + \frac{1}{2M}\int_{[0,1]^2} \big(K(l,l') - K_n^*(l,l')\big)^2 dl dl' \\
&\qquad + \xi_n \mathcal{I}_n^{\text{reg}}[K] - \xi_n \mathcal{I}_n^{\text{reg}}[K_n^*] \\
&\overset{(b)}{\geq} \text{Tr}(T_\nabla(T_K - T_{K_n^*})) + \xi_n \text{Tr}(V^*(T_K - T_{K_n^*})) + \frac{1}{2M}\int_{[0,1]^2} \big(K(l,l') - K_n^*(l,l')\big)^2 dl dl' \\
&\overset{(c)}{\geq} \frac{1}{2M}\int_{[0,1]^2} \big(K(l,l') - K_n^*(l,l')\big)^2 dl dl'.
\end{aligned}$$

To obtain this, in (a) we substituted in the bound on $\mathcal{I}_n[K] - \mathcal{I}_n[K']$ stated above. In (b), we used the isometry between the trace inner product on operators and the corresponding inner product of the kernels, and the KKT conditions stating the existence of a bounded operator V^* for which $\text{Tr}(V^* T_{K_n^*}) = \mathcal{I}_n^{\text{reg}}[K_n^*]$ and $\|V^*\|_{\text{op}} \leq 1$; the latter property consequently implies that $\mathcal{I}_n^{\text{reg}}[K] \geq \text{Tr}(V^* K)$ by the variational formulation of the trace. In (c), we then use the fact that K_n^* is optimal provided that $\text{Tr}((T_\nabla + \xi_n V^*)(T_K - T_{K_n^*})) \geq 0$. □

E Proof of additional theorems from Section 3.1

In this section, we write $\mu_i(K)$ for either the i-th largest eigenvalue of a symmetric matrix K, or the i-th largest eigenvalue of a self-adjoint operator with kernel K (as introduced in the beginning of Appendix D. We write $\sigma_i(K)$ for the corresponding singular values; recall that for a matrix $K \in \mathbb{R}^{n \times d}$, we have that $\sigma_r(K)^2 = \mu_r(KK^T)$ for any $r \leq \min\{n,d\}$, and that for a self-adjoint positive definite matrix or operator K, we have that $\sigma_r(K) = \mu_r(K)$ for all r.

Before proving Theorems 3 and 4, we require a brief lemma.

Lemma 9. *Let $K : [0,1]^2 \to \mathbb{R}$ be the kernel of a symmetric, positive operator which is either a) piecewise constant on a partition $\mathcal{Q} \times \mathcal{Q}$ where $\mathcal{Q}$ is a partition of $[0,1]$ of size κ, or b) continuous. Suppose moreover that K has rank exactly equal to r, where*

$$K(x,y) = \sum_{i=1}^r \psi_i(x)\psi_i(y) \tag{55}$$

for some non-zero, orthogonal functions $\phi_i : [0,1] \to \mathbb{R}$ which are piecewise continuous. Then if λ_i are i.i.d $\text{Unif}([0,1])$ and we define the random matrix $(K_\lambda)_{ij} := K(\lambda_i, \lambda_j)$, then K_λ is of rank $\leq r$, and with asymptotic probability 1 as $n \to \infty$, has rank exactly equal to r.

Proof of Lemma 9. Note that if we write $\Psi_{ij} = \psi_j(\lambda_i) \in \mathbb{R}^{n\times r}$, then as $K_\lambda = \Psi\Psi^T$, we know that the rank of K_λ must be less than or equal to r. For the second part, we note that under the given conditions, we can apply Corollary 5.5 of [36] to the matrix $n^{-1}K_\lambda$; under a), the diagonal summability condition needed follows trivially, and under b), Mercer's theorem gives the diagonal summability condition needed, with the other conditions being satisfied as a result of K being finite rank. Consequently we have that

$$\mu_r(n^{-1}K_\lambda) = \mu_r(K) + O_p(n^{-1/2}) > \frac{1}{2}\mu_r(K) \text{ with probability} \to 1.$$

In particular, with asymptotic probability 1, $n^{-1}K_\lambda$ is of full rank, and therefore so is K_λ. □

Proof of Theorem 3. To save on notation, we write K for K_n^*, K_λ for the matrix $(K(\lambda_i, \lambda_j))_{ij}$, and $\phi_i(l)$ for the $\phi_{n,i}(l)$. We note that Lemma 9 gives the guarantee that K_λ is asymptotically of exact rank r. Writing $\Psi_\lambda \in \mathbb{R}^{n\times r}$ for the matrix $(\phi_j(\lambda_i))$ for $i \in [n]$ and $j \in [r]$, the same argument in Lemma 9 guarantees that the singular value $\sigma_r(n^{-1/2}\Psi_\lambda)^2 = \mu_r(n^{-1}K_\lambda) \geq \frac{1}{2}\mu_r(K) > 0$ with asymptotic probability 1, and therefore we can work on an event where the r-th highest singular value of $n^{-1/2}\Psi_\lambda$ is uniformly bounded away from zero.

With this, we can now apply Lemma 5.4 of [60], which states that for any matrices $U, V \in \mathbb{R}^{n\times r}$, we have that

$$\min_{Q\in O(r)} \|U - VQ\|_F^2 \leq \frac{1}{2(\sqrt{2}-1)\sigma_r^2(V)}\|UU^T - VV^T\|_F^2,$$

where $\sigma_d(V)$ is the d-th largest singular value of the matrix V. We recall that $\sigma_r(V)^2 = \mu_r(VV^T)$. Applying this to $U = n^{-1/2}\boldsymbol{\omega}_n$ and $V = n^{-1/2}\Psi_\lambda$, followed by the above remark, gives the desired result. □

Proof of Theorem 4. For this, we begin by noting that as $\tilde{G}$ is defined to be a best rank r approximation to the matrix G, $n^{-1}\tilde{G}$ is a best rank r approximation to the matrix $n^{-1}G$, and consequently we have that

$$n^{-2}\|\tilde{G} - G\|_F^2 = \|n^{-1}\tilde{G} - n^{-1}G\|_F^2 = \sum_{i=r+1}^{d} \mu_i(n^{-1}G)^2 \tag{56}$$

by the Eckart–Young–Mirsky theorem. To proceed, we then recall that as $G_{ij} = \langle\widehat{\omega}_i, \widehat{\omega}_j\rangle$, and moreover we have that $\mu_i(K_\lambda) = 0$ for $i \geq r+1$ we have that

$$\begin{aligned}\sum_{i=r+1}^{d} \mu_i(n^{-1}G)^2 &= \sum_{i=r+1}^{d} \big(\mu_i(n^{-1}G) - \mu_i(n^{-1}K_\lambda)\big)^2 \\ &\leq \sum_{i=1}^{d} \big(\mu_i(n^{-1}G) - \mu_i(n^{-1}K_\lambda)\big)^2 \leq \|n^{-1}G - n^{-1}K_\lambda\|_F^2 = o_p(1)\end{aligned} \tag{57}$$

where the last inequality follows by the Weilandt-Hoffman inequality [30], giving the first part of the theorem statement. The second part of the theorem statement then follows by applying the proof of Theorem 3 to the matrix $\tilde{\Omega}$, noting that

$$n^{-2}\|\tilde{G} - K_\lambda\|_F^2 \leq n^{-2}\|\tilde{G} - G\|_F^2 + n^{-2}\|G - K_\lambda\|_F^2 \leq 2n^{-2}\|G - K_\lambda\|_F^2 = o_p(1)$$

by the triangle inequality and by combining (56) and (57) together. □

Before proving Theorem 5, we require a lemma about the eigenfunctions of an operator whose kernel is Hölder continuous.

Lemma 10. *Suppose that $K : [0,1]^2 \to \mathbb{R}$ is symmetric and Hölder$([0,1]^2, \beta, L)$ continuous. Then the eigenfunctions of the associated operator T_K are Hölder$([0,1], \beta, L)$ continuous, and moreover are uniformly bounded in $L^\infty([0,1])$.*

Proof of Lemma 10. We begin by noting that for any function $f \in L^2([0,1])$, we have that

$$\begin{aligned}\left|T_K[f](x) - T_K[f](y)\right| &\leq \int_0^1 |K(x,z) - K(y,z)| \cdot |f(z)|\, dx \\ &\leq L\|f\|_1 |x-y|^\beta \leq L\|f\|_2 |x-y|^\beta,\end{aligned}$$

and therefore the image of the unit ball $\|f\|_2 = 1$ consists of Hölder$([0,1], \beta, L)$ continuous functions; consequently, so are the eigenvectors. Moreover, we note that the image of such a ball gives functions which are uniformly bounded in $L^\infty([0,1])$; indeed, writing $g = T_K[f]$, and picking any $x \in [0,1]$, we have that

$$|g(x)| \leq |g(x) - g(y)| + |g(y)| \text{ for all } y \in [0,1]$$

and therefore by integrating against y we end up with

$$|g(x)| \leq \int_0^1 |g(x) - g(y)|\, dy + \int_0^1 |g(y)|\, dy \leq L \int_0^1 |x-y|^\beta\, dy + \|g\|_1 \leq L + 1$$

as $\|g\|_1 \leq \|g\|_2 = 1$, and $|x-y|^\beta \leq 1$ for all $x, y \in [0,1]$. □

Proof of Theorem 5. Without loss of generality, suppose that $c_1 = c_2 = 1$; otherwise, we can just rescale the regularization constant ξ so that, up to constant scaling, the objective is the same as one with $c_1 = c_2 = 1$. Now, recall that by the spectral theorem for self-adjoint operators and Lemma 10, we can write

$$T_W[f] = \sum_{i=1}^\infty \mu_i(W)\langle f, \phi_i\rangle \phi_i \qquad \text{and} \qquad W(l,l') = \sum_{i=1}^\infty \mu_i(W)\phi_i(l)\phi_i(l')$$

where the latter sum converges in L^2, the $\mu_i(W)$ are sorted in monotone decreasing absolute value, the $(\phi_i)_{i\geq 1}$ are orthonormal eigenfunctions which are Hölder$([0,1], \beta, L)$ and uniformly bounded in $L^\infty([0,1])$. We now want to study the minimizer of the function

$$\|T_W - T_L\|_{\text{HS}}^2 + \xi\|T_L\|_1$$

over all positive kernels L, where we have phrased the problem entirely in terms of the associated operators. To do so, we begin by writing

$$T_L = T_{L^\parallel} + T_{L^\perp} \qquad \text{where } L^\parallel(x,y) = \sum_{n=1}^\infty \mu_i \phi_n(x)\phi_n(y),$$

for some $\mu_i \geq 0$, where $L^\perp$ is symmetric, positive and orthogonal to $L^\parallel$ in that $L^\perp[\phi] = 0$ for any $\phi \in \text{cl}\{\text{span}(\phi_1, \phi_2, \ldots)\}$. We can then argue that any minimizer L must have $L^\perp = 0$. Indeed, we have that by orthogonality of $T_{L^\perp}$ to both T_W and $T_{L^\parallel}$, we get the decomposition

$$\|T_W - T_L\|_{\text{HS}}^2 = \|T_W - T_{L^\parallel}\|_{\text{HS}}^2 + \|T_{L^\perp}\|_{\text{HS}}^2$$

and so $\|T_W - T_{L^\parallel}\|_{\text{HS}}^2 \leq \|T_W - T_L\|_{\text{HS}}^2$ with equality if and only if $T_{L^\perp} = 0$; and moreover $\|T_{L^\parallel}\|_1 \leq \|T_L\|_1$. As $T_{L^\perp} = 0$, we can then show that the objective function equals

$$\sum_{i=1}^\infty (\mu_i - \mu_i(W))^2 + \xi \sum_{i=1}^\infty \mu_i.$$

To minimize this, we note that we can minimize each term in the sum over $\mu_i \geq 0$ by taking $\widehat{\mu}_i = (\mu_i(W) - \xi)_+$. In particular, as the eigenvalues of W decay as $O(i^{-(1/2+\beta)})$ [51], it follows that for $i \geq N$ where $N = O(\xi^{-2/(2+2\beta)})$, we have that $\widehat{\mu}_i = 0$. Consequently, we get that

$$\widehat{L}(x,y) = \sum_{i=1}^N \widehat{\mu}_i \phi_i(x)\phi_i(y)$$

is the minimizing positive kernel. We now note that as the ϕ_i are uniformly bounded in $L^\infty([0,1])$, and the $\widehat{\mu}_i$ are bounded above also, we can argue that $\widehat{L}$ will belong to some set $\mathcal{Z}_d^{\geq 0}(A)$ for some A sufficiently large and any $d \geq N$, and consequently $\widehat{L} \in \mathcal{Z}^{\geq 0}$ also. In particular, this means that $\widehat{L}$ is the minimizer of the objective function over the set $\mathcal{Z}^{\geq 0}$. Finally, we then note that as the eigenfunctions are Hölder and we have a finite sum of terms of the form $\widehat{\mu}_i\phi_i(x)\phi_i(y)$, this plus the boundedness of the eigenfunctions will imply that $\widehat{L}$ is Hölder of exponent β also. □

F Proof of results in Section 3.2

In this section, given triangular arrays (X_{ni}) and (Y_{ni}) for $i \in I_n$, $n \geq 1$, we use the notation $X_{n,i} = (1 + O_p(r_n))Y_{n,i}$ to be equivalent to saying that $\max_{i \in I_n} |X_{ni}/Y_{ni} - 1| = O_p(r_n)$. Before giving the proofs of Lemmas 1, 2 and 3, we require the following result.

Lemma 11. *Suppose that Assumption 2 holds. Let $g : [0,1]^2 \to [0,\infty]$ be a bounded measurable function which is bounded away from zero, and define*

$$T_{n,i} = \frac{1}{n-1} \sum_{j \in [n] \setminus \{i\}} a_{ij} g(\lambda_j), \qquad \textit{so} \qquad \mathbb{E}[T_{n,i} \,|\, \lambda_i] = \rho_n \int_0^1 W(\lambda_i, y) g(y)\, dy.$$

Then for all $t \geq 0$ we have that

$$\mathbb{P}\Big(\Big| \frac{T_{n,i}}{\mathbb{E}[T_{n,i} \,|\, \lambda_i]} - 1 \Big| \geq t \,|\, \lambda_i \Big) \leq 2 \exp\Big(\frac{-n \mathbb{E}[T_{n,i} \,|\, \lambda_i] t^2}{8 \|g\|_\infty (1+t)} \Big)$$

and whence that $T_{n,i} = \mathbb{E}[T_{n,i} \,|, \lambda_i](1 + O_p((\log n / n\rho_n)^{1/2}))$. Similarly, if we write

$$\widetilde{T}_{n,i} = \frac{1}{n-1} \sum_{j \in [n] \setminus \{i\}} (1 - a_{ij}) g(\lambda_j), \textit{ so } \mathbb{E}[\widetilde{T}_{n,i} \,|\, \lambda_i] = \int_0^1 (1 - \rho_n W(\lambda_i, y)) g(y)\, dy,$$

then $\widetilde{T}_{n,i} = \mathbb{E}[\widetilde{T}_{n,i} \,|, \lambda_i](1 + O_p((\log n/n)^{1/2}))$.

To prove this result we use the method of exchangeable pairs to derive a concentration inequality. Assuming that (X, X') is an exchangeable pair of random variables, and f is a measurable function with $\mathbb{E}[f(X)] = 0$, if we have an anti-symmetric function $F(X, X')$ satisfying

$$\mathbb{E}[F(X,X')|X] = f(X), \qquad v(X) := \frac{1}{2}\mathbb{E}\big[\big| \{f(X) - f(X')\} F(X,X') \big| \,|\, X \big] \leq B f(X) + C,$$

for some constants $B, C \geq 0$, then we get the concentration inequality [17, Theorem 3.9]

$$\mathbb{P}\big(|f(X)| \geq t\big) \leq 2 \exp\Big(- \frac{t^2}{2C + 2Bt} \Big). \tag{58}$$

Proof of Lemma 11. We begin by noting that as g is assumed to be bounded away from zero, and by Assumption 2 we assume that W is also bounded away from zero, there exists a constant $c > 0$ such that $\mathbb{E}[T_{n,i}|\lambda_i] \geq c\rho_n > 0$ for all $i \in [n], n \geq 1$. To derive the given bounds, we will use the method of exchangeable pairs, working conditional on the λ_i at first in order to derive a concentration inequality. By then using the above lower bound on $\mathbb{E}[T_{n,i}|\lambda_i]$, we will be able to obtain a bound which holds unconditionally, and consequently get the claimed bound on $T_{n,i}$ holding uniformly across all the vertices.

To begin, let $\boldsymbol{A}_{n,i}$ denote the i-th row of the adjacency matrix $\boldsymbol{A}_n$, and $\boldsymbol{\lambda}_{n,-i} := (\lambda_j)_{j \neq i}$. We construct an exchangeable pair $\big((\boldsymbol{\lambda}_{n,-i}, \boldsymbol{A}_{n,i}), (\tilde{\boldsymbol{\lambda}}_{n,-i}, \tilde{\boldsymbol{A}}_{n,i})\big)$ as follows: we select an index J uniformly from $[n] \setminus \{i\}$, redraw $\tilde{\lambda}_J \sim U[0,1]$ and $\tilde{a}_{iJ} \sim \mathrm{Bern}(W_n(\lambda_i, \tilde{\lambda}_J))$ but otherwise we keep the other entries of $\tilde{\boldsymbol{\lambda}}_n$ and $\tilde{\boldsymbol{A}}_{n,i}$ the same. With this, note that

$$\frac{1}{\mathbb{E}[T_{n,i} \,|\, \lambda_i]} \mathbb{E}\Big[\sum_{j \in [n] \setminus \{i\}} a_{ij} g(\lambda_j) - \sum_{j \in [n] \setminus \{i\}} \tilde{a}_{ij} g(\tilde{\lambda}_j) \,\Big|\, \lambda_i, \boldsymbol{\lambda}_{n,-i}, \boldsymbol{A}_{n,i} \Big] = \frac{T_{n,i}}{(n-1)\mathbb{E}[T_{n,i} \,|\, \lambda_i]} - 1,$$

and the associated variance term is of the form

$$\begin{aligned} v\big(\boldsymbol{\lambda}_{n,-i}, \boldsymbol{A}_{n,i}\big) &= \frac{1}{(n-1)\mathbb{E}[T \,|\, \lambda_i]^2} \mathbb{E}\Big[\Big(\sum_{j \in [n] \setminus \{i\}} \{a_{ij} g(\lambda_j) - \tilde{a}_{ij} g(\tilde{\lambda}_j)\} \Big)^2 \,\Big|\, \lambda_i, \boldsymbol{\lambda}_{n,-i}, \boldsymbol{A}_{n,i} \Big] \\ &= \frac{1}{(n-1)^2 \mathbb{E}[T \,|\, \lambda_i]^2} \sum_{j \in [n] \setminus \{i\}} \mathbb{E}\Big[\big\{ a_{ij} g(\lambda_j) - a'_{ij} g(\lambda'_j) \big\}^2 \,\Big|\, \lambda_i, \boldsymbol{\lambda}_{n,-i}, \boldsymbol{A}_{n,i} \Big], \end{aligned}$$

where $(a'_{ij})_{ij}$ and $(\lambda'_i)_{i\geq 1}$ are independent copies of $(a_{ij})_{ij}$ and $(\lambda_i)_{i\geq 1}$. To bound this last quantity, we write

$$\begin{aligned}
&\mathbb{E}\Big[\big\{a_{ij}g(\lambda_j) - a'_{ij}g(\lambda'_j)\big\}^2 \,\Big|\, \lambda_i, \boldsymbol{\lambda}_{n,-i}, \boldsymbol{A}_{n,i}\Big] \\
&\qquad = \|g\|_\infty^2 \mathbb{E}\Big[\Big\{a_{ij}\frac{g(\lambda_j)}{\|g\|_\infty} - a'_{ij}\frac{g(\lambda'_j)}{\|g\|_\infty}\Big\}^2 \,\Big|\, \lambda_i, \boldsymbol{\lambda}_{n,-i}, \boldsymbol{A}_{n,i}\Big] \\
&\qquad \leq 2\|g\|_\infty \mathbb{E}\Big[a_{ij}g(\lambda_j) + a'_{ij}g(\lambda'_j) \,\Big|\, \lambda_i, \boldsymbol{\lambda}_{n,-i}, \boldsymbol{A}_{n,i}\Big],
\end{aligned}$$

where we used the fact the inequalities $(a-b)^2 \leq 2(a^2+b^2) \leq 2(a+b)$ for $a, b \in [0,1]$ to obtain the last line. It therefore follows that

$$v\big(\boldsymbol{\lambda}_{n,-i}, \boldsymbol{A}_{n,i}\big) \leq \frac{2\|g\|_\infty}{(n-1)\mathbb{E}[T_{n,i}|\lambda_i]}\Big(\frac{T_{n,i}}{(n-1)\mathbb{E}[T_{n,i}|\lambda_i]} + 1\Big)$$

from which we can apply the inequality stated in (58) to get the stated concentration inequality. As $\mathbb{E}[T_{n,i}]|\lambda_i] \geq c\rho_n$, we can conclude that

$$\mathbb{P}\Big(\Big|\frac{T_{n,i}}{\mathbb{E}[T_{n,i} \,|\, \lambda_i]} - 1\Big| \geq t\Big) \leq 2\exp\Big(\frac{-cn\rho_n t^2}{8\|g\|_\infty(1+t)}\Big)$$

for all $i \in [n]$, from which taking a union bound allows us to conclude that $T_{n,i} = \mathbb{E}[T_{n,i} \,|\, \lambda_i](1 + O_p((\log(n)/n\rho_n)^{1/2}))$. The same style of argument gives the claimed result when $a_{ij} \to 1 - a_{ij}$, noting that in this case one can instead argue that $\mathbb{E}[\tilde{T}_{n,i} \,|\, \lambda_i] \geq c'$ for some constant $c' > 0$ for all $i \in [n]$, $n \geq 1$. □

Proof of Lemma 1. We note that a vertex i is sampled with probability k/n, and any pair of vertices (i,j) is sampled with probability $k(k-1)/n(n-1)$, so the claimed result follows immediately. □

Proof of Lemma 2. The formulae for $f_n(l,l',1)$ and $f_n(l,l',0)$ are given in Proposition 72 of [21]. It remains to derive the formula for $\tilde{g}_n(\lambda_i)$. For convenience, we denote $\tilde{s}_n = (\log(n)/n\rho_n)^{1/2}$. To continue, we note that in the proof of Proposition 72 of [21], it is shown that

$$\begin{aligned}
\mathbb{P}\big(u \in \mathcal{V}(S_0(\mathcal{G}_n)) \,|\, \mathcal{G}_n\big) &= \frac{2kW(\lambda_u,\cdot)}{\mathcal{E}_W n}\big(1 + O_p\big(\tilde{s}_n\big)\big), \\
\mathbb{P}\big(B(l, \mathrm{Ug}_\alpha(u \,|\, \mathcal{G}_n) \geq 1 \,|\, \mathcal{G}_n\big) &= \frac{lW(\lambda_u,\cdot)^\alpha}{n\mathcal{E}_W(\alpha)}\big(1 + O_p(\tilde{s}_n))\big), \\
\mathbb{P}\big(u, v \in \mathcal{V}(S_0(\mathcal{G}_n)) \,|\, \mathcal{G}_n &= \Big(\frac{2ka_{uv}}{n^2\rho_n\mathcal{E}_W} + \frac{4k(k-1)W(\lambda_u,\cdot)W(\lambda_v,\cdot)}{n^2\mathcal{E}_W^2}\Big)\cdot\big(1 + O_p(\tilde{s}_n)\big),
\end{aligned}$$

and as a particular consequence, it therefore follows that

$$\mathbb{P}\big(u \in \mathcal{V}(S_0(\mathcal{G}_n)) \,\big|\, v \in \mathcal{V}(S_0(\mathcal{G}_n)), \mathcal{G}_n\big) = \Big(\frac{a_{uv}}{n\rho_n W(\lambda_v,\cdot)} + \frac{2(k-1)W(\lambda_u,\cdot)}{n\mathcal{E}_W}\Big)\cdot\big(1 + O_p(\tilde{s}_n)\big).$$

To begin in finding the formula for $\tilde{g}_n(\lambda_i)$, we note that

$$\mathbb{P}\big(u \in \mathcal{V}(S(\mathcal{G}_n)) \,|\, \mathcal{G}_n\big) = \mathbb{P}\big(\mathcal{V}(u \in S_0(\mathcal{G}_n)) \,|\, \mathcal{G}_n\big) + \mathbb{P}\big(u \in \mathcal{V}(S_{ns}(\mathcal{G}_n) \setminus S_0(\mathcal{G}_n)) \,|\, \mathcal{G}_n\big),$$

where the first term is given as above. The second term corresponds to the probability that the vertex arises only through the negative sampling process, and so

$$\mathbb{P}\big(u \in \mathcal{V}(S_{ns}(\mathcal{G}_n) \setminus S_0(\mathcal{G}_n)) \,|\, \mathcal{G}_n\big) = \mathbb{P}\Big(\bigcup_{v \in \mathcal{V}_n \setminus \{u\}} A_v \,\Big|\, \mathcal{G}_n\Big)$$

where $A_v = \{v \in \mathcal{V}(S_0(\mathcal{G}_n)), u \text{ selected via negative sampling from } v\}$. We then have that

$$\begin{aligned}
\Big|\mathbb{P}\Big(\bigcup_{v \in \mathcal{V}_n \setminus \{u\}} A_v \,\Big|\, \mathcal{G}_n\Big) - \sum_{v \in \mathcal{V}_n \setminus \{u\}} \mathbb{P}(A_v \,|\, \mathcal{G}_n)\Big| &\leq \frac{1}{2}\sum_{\substack{v,v' \in \mathcal{V}_n \setminus \{u\} \\ v' \neq v}} \mathbb{P}\big(A_v \cap A_{v'} \,|\, \mathcal{G}_n\big) \\
&\leq \sum_{v \in \mathcal{V}_n \setminus \{u\}} \mathbb{P}\big(A_v \,|\, \mathcal{G}_n\big) \cdot \max_{v' \in \mathcal{V}_n \setminus \{u\}} \sum_{v \in \mathcal{V}_n \setminus \{v', u\}} \mathbb{P}\big(A_v \,|\, A_{v'}, \mathcal{G}_n\big).
\end{aligned}$$

We begin by finding the asymptotic form of $\sum_{v\in\mathcal{V}_n\setminus\{u\}}\mathbb{P}(A_v\,|\,\mathcal{G}_n)$, where we find that

$$\begin{aligned}\sum_{v\in\mathcal{V}_n\setminus\{u\}}\mathbb{P}\big(A_v\,|\,\mathcal{G}_n\big) &= \sum_{v\neq u}\mathbb{P}\big(v\in\mathcal{V}(S_0(\mathcal{G}_n))\big)\mathbb{P}\big(B(l,\mathrm{Ug}_\alpha(u\,|\,\mathcal{G}_n)\geq 1\,|\,\mathcal{G}_n\big)(1-a_{uv})\\ &= \big(1+O_p\big(\tilde{s}_n\big)\big)\cdot\frac{2klW(\lambda_u,\cdot)^\alpha}{n\mathcal{E}_W(\alpha)\mathcal{E}_W}\cdot\frac{1}{n}\sum_{v\in\mathcal{V}_n\setminus\{u\}}(1-a_{uv})W(\lambda_v,\cdot)\\ &= \big(1+O_p\big(\tilde{s}_n\big)\big)\frac{2klW(\lambda_u,\cdot)^\alpha}{n\mathcal{E}_W(\alpha)\mathcal{E}_W}\cdot\int_0^1(1-\rho_nW(\lambda_u,y))W(y,\cdot)\,dy,\end{aligned}$$

where we have used the formulae quoted at the beginning of the proof and Lemma 11. It remains to examine the term

$$\max_{v'\in\mathcal{V}_n\setminus\{u\}}\sum_{v\in\mathcal{V}_n\setminus\{v',u\}}\mathbb{P}\big(A_v\,|\,A_{v'},\mathcal{G}_n\big).$$

To do so, note that we can write

$$\begin{aligned}&\mathbb{P}\big(A_v\cap A_{v'}\,|\,\mathcal{G}_n\big)\\ &\qquad=\mathbb{P}\big(v,v'\in\mathcal{V}(S_0(\mathcal{G}_n))\,|\,\mathcal{G}_n\big)\mathbb{P}\big(B(l,\mathrm{Ug}_\alpha(u\,|\,\mathcal{G}_n))\geq 1\,|\,\mathcal{G}_n\big)^2(1-a_{uv})(1-a_{uv'})\end{aligned}$$

and so

$$\begin{aligned}&\mathbb{P}\big(A_v\,|\,A_{v'},\mathcal{G}_n\big)\\ &\qquad=\mathbb{P}\big(v\in\mathcal{V}(S_0(\mathcal{G}_n))\,|\,v'\in\mathcal{V}(S_0(\mathcal{G}_n)),\mathcal{G}_n\big)\mathbb{P}\big(B(l,\mathrm{Ug}_\alpha(u\,|\,\mathcal{G}_n))\geq 1\,|\,\mathcal{G}_n\big)(1-a_{uv}).\end{aligned}$$

It therefore follows that, using the results stated at the beginning of the proof,

$$\begin{aligned}&\sum_{v\in\mathcal{V}_n\setminus\{v',u\}}\mathbb{P}\big(A_v\,|\,A_{v'},\mathcal{G}_n\big)\\ &=(1+O_p(\tilde{s}_n))\frac{lW(\lambda_u,\cdot)^\alpha}{n\mathcal{E}_W(\alpha)}\sum_{v\in\mathcal{V}_n\setminus\{v',u\}}\mathbb{P}\big(v\in\mathcal{V}(S_0(\mathcal{G}_n))\,|\,v'\in\mathcal{V}(S_0(\mathcal{G}_n)),\mathcal{G}_n\big)(1-a_{uv})\\ &=(1+O_p(\tilde{s}_n))\frac{lW(\lambda_u,\cdot)^\alpha}{n^2\mathcal{E}_W(\alpha)}\sum_{v\in\mathcal{V}_n\setminus\{v',u\}}(1-a_{uv})\cdot\big(\frac{a_{vv'}}{\rho_nW(\lambda_{v'},\cdot)}+2(k-1)\mathcal{E}_W^{-1}W(\lambda_v,\cdot)\big)\\ &=O_p(n^{-1})\end{aligned}$$

uniformly across all v',u, and therefore

$$\max_{v'\in\mathcal{V}_n\setminus\{u\}}\sum_{v\in\mathcal{V}_n\setminus\{v',u\}}\mathbb{P}\big(A_v\,|\,A_{v'},\mathcal{G}_n\big)=O_p(n^{-1}).$$

Combining all of the above together then gives that

$$\begin{aligned}&\mathbb{P}\big(u\in\mathcal{V}(S(\mathcal{G}_n))\,|\,\mathcal{G}_n\big)\\ &\qquad=\frac{2k}{n\mathcal{E}_W}\Big(W(\lambda_u,\cdot)+\frac{lW(\lambda_u,\cdot)^\alpha}{\mathcal{E}_W(\alpha)}\cdot\int_0^1(1-\rho_nW(\lambda_u,y))W(y,\cdot)\,dy\Big)\big(1+O_p\big(\tilde{s}_n\big)\big)\end{aligned}$$

and so we get the stated formula for g_n with $s_n=\tilde{s}_n$. □

Proof of Lemma 3. The formulae for $f_n(l,l',1)$ and $f_n(l,l',0)$ are given in Proposition 74 of [21]. We also note that within the proof of Proposition 74 of [21], we have that

$$\begin{aligned}\mathbb{P}\big(u\in\mathcal{V}(S_0(\mathcal{G}_n))\,|\,\mathcal{G}_n\big)&=\frac{kW(\lambda_u,\cdot)}{n\mathcal{E}_W}\big(1+O_p\big(\tilde{s}_n\big)\big),\\ \mathbb{P}\big(\tilde{v}_i=u\,|\,\mathcal{G}_n\big)&=\frac{W(\lambda_u,\cdot)}{n\mathcal{E}_W}\big(1+O_p(\tilde{s}_n)\big),\\ \mathbb{P}\big(u\text{ selected via negative sampling from }v\,|\,\mathcal{G}_n\big)&=\frac{lW(\lambda_u,\cdot)^\alpha(1-a_{uv})}{n\mathcal{E}_W(\alpha)}\big(1+O_p(\tilde{s}_n)\big),\end{aligned}$$

where we again write $\tilde{s}_n = (\log(n)/n\rho_n)^{1/2}$. To derive the corresponding formula for $\tilde{g}_n(l)$, we begin by noting

$$\mathbb{P}\big(u \in \mathcal{V}(S(\mathcal{G}_n)) \,|\, \mathcal{G}_n\big) = \mathbb{P}\big(u \in \mathcal{V}(S_0(\mathcal{G}_n)) \,|\, \mathcal{G}_n\big) + \mathbb{P}\big(u \in \mathcal{V}(S_{ns}(\mathcal{G}_n) \setminus S_0(\mathcal{G}_n)) \,|\, \mathcal{G}_n\big).$$

The first term is given above, so we focus on the second. Denoting $A_i(u) = \{\tilde{v}_i = u\}$ for $i \leq k+1$ and $u \in \mathcal{V}_n$, and $B_i(v|u) = \{v \text{ selected via negative sampling from } u\}$, we know that

$$\mathbb{P}\big(u \in \mathcal{V}(S(\mathcal{G}_n) \setminus S_0(\mathcal{G}_n)) \,|\, \mathcal{G}_n\big) = \mathbb{P}\Big(\bigcup_{i=1}^{k+1} \bigcup_{v \in [n]\setminus\{u\}} A_i(v) \cap B_i(u|v) \,|\, \mathcal{G}_n\Big).$$

Letting $C_i = \cup_{v\in[n]\setminus\{u\}}\{A_i(v) \cap B_i(u|v)\}$, we note that

$$\Big(\sum_{i=1}^{k+1} 1[C_i]\Big) - 1\Big[\bigcup_{j=1}^{k+1} C_j\Big] = \sum_{i=1}^{k} 1\Big[C_i \cap \cup_{j>i} C_j\Big],$$

and moreover that as the $A_i(v)$ are disjoint across all $v \in \mathcal{V}_n$ for each i fixed, we have that

$$\sum_{i=1}^{k+1} \mathbb{P}\Big(\bigcup_{v\in[n]\setminus\{u\}} A_i(v) \cap B_i(u|v) \,|\, \mathcal{G}_n\Big) = \sum_{i=1}^{k+1} \sum_{v\in[n]\setminus\{u\}} \mathbb{P}\Big(A_i(v) \cap B_i(u|v) \,|\, \mathcal{G}_n\Big).$$

Combining the above two facts therefore gives

$$\begin{aligned}
&\Big|\mathbb{P}\Big(\bigcup_{i=1}^{k+1} \bigcup_{v\in[n]\setminus\{u\}} A_i(v) \cap B_i(u|v) \,|\, \mathcal{G}_n\Big) - \sum_{i=1}^{k+1} \sum_{v\in[n]\setminus\{u\}} \mathbb{P}\big(A_i(v) \cap B_i(u|v) \,|\, \mathcal{G}_n\big)\Big| \\
&\qquad \leq \sum_{i=1}^{k} \mathbb{P}\Big(\Big\{\bigcup_{v\in[n]\setminus\{u\}} A_i(v) \cap B_i(u|v)\Big\} \cap \Big\{\bigcup_{j>i} \bigcup_{v'\in[n]\setminus\{u\}} A_j(v') \cap B_j(u|v')\Big\} \,|\, \mathcal{G}_n\Big) \\
&\qquad \leq \sum_{i=1}^{k} \sum_{j>i} \sum_{v,v'\in[n]\setminus\{u\}} \mathbb{P}\big(A_j(v') \cap B_j(u|v') \cap A_i(v) \cap B_i(u|v) \,|\, \mathcal{G}_n\big).
\end{aligned}$$

To handle the intersection probabilities, we note that we can write (using the above formulae), for indices $i < j$ and v, v', that

$$\begin{aligned}
&\mathbb{P}\big(A_i(v) \cap B_i(u|v) \cap A_j(v') \cap B_j(u|v') \,|\, \mathcal{G}_n\big) \\
&\qquad = \big(1 + O_p(\tilde{s}_n)\big) \cdot \frac{l^2 W(\lambda_u,\cdot)^{2\alpha}}{n^2 \mathcal{E}_W(\alpha)^2} \mathbb{P}\big(A_j(v') \,|\, A_i(v)\mathcal{G}_n\big)\mathbb{P}\big(A_i(v) \,|\, \mathcal{G}_n\big)(1-a_{uv})(1-a_{uv'}) \\
&\qquad = \big(1 + O_p(\tilde{s}_n)\big) \cdot \frac{l^2 W(\lambda_u,\cdot)^{2\alpha} W(\lambda_v,\cdot)}{n^3 \mathcal{E}_W(\alpha)^2 \mathcal{E}_W} \mathbb{P}\big(A_j(v') \,|\, A_i(v), \mathcal{G}_n\big)(1-a_{uv})(1-a_{uv'}).
\end{aligned}$$

Write E_n for the number of edges in $\mathcal{G}_n$ and $\deg_n(u)$ for the degree of the vertex u in $\mathcal{G}_n$; then by Proposition 61 of [21] we have that

$$\max_{u\in[n]} \frac{1}{\deg_n(u)} = O_p((n\rho_n)^{-1}), \qquad \deg_n(u) = n\rho_n W(\lambda_u,\cdot)(1 + O_p(\tilde{s}_n)),$$

and we note that by Assumption 3 that $W(\lambda_u,\cdot)$ is bounded below away from zero (and above by one) uniformly over all $\lambda_u \in [0,1]$. To handle the $\mathbb{P}(A_j(v') \,|\, A_i(v),\, \mathcal{G}_n)$ term, we note that by stationarity of the random walk and the Markov property that, when $j > i+1$

$$\begin{aligned}
\mathbb{P}(A_j(v') \,|\, A_i(v),\, \mathcal{G}_n) &= \mathbb{P}(\tilde{v}_{j-i+1} = v' \,|\, \tilde{v}_1 = v,\, \mathcal{G}_n) \\
&= \sum_{u\,:\,a_{uv}=1} \mathbb{P}(\tilde{v}_{j-i+1} = v' \,|\, \tilde{v}_2 = u, \mathcal{G}_n)\mathbb{P}(\tilde{v}_2 = u \,|\, \tilde{v}_1 = v, \mathcal{G}_n) \\
&= \sum_{u\,:\,a_{uv}=1} \frac{2E_n}{\deg_n(u)\deg_n(v)} \mathbb{P}(\tilde{v}_{j-i+1} = v' \,|\, \tilde{v}_2 = u, \mathcal{G}_n)\mathbb{P}(\tilde{v}_2 = u \,|\, \mathcal{G}_n)
\end{aligned}$$

$$\leq \sum_{u\in[n]} \frac{2E_n}{\deg_n(u)\deg_n(v)} \mathbb{P}(\tilde{v}_{j-i+1} = v' \,|\, \tilde{v}_2 = u, \mathcal{G}_n)\mathbb{P}(\tilde{v}_2 = u \,|\, \mathcal{G}_n)$$
$$\leq \frac{\deg_n(v')}{\deg_n(v)} \cdot \max_{u\in[n]} \frac{1}{\deg_n(u)} = O_p((n\rho_n)^{-1}),$$

and when $j = i+1$ we have

$$\mathbb{P}(A_{i+1}(v') \,|\, A_i(v),\, \mathcal{G}_n) = \frac{a_{vv'}}{\deg_n(v)} = O_p((n\rho_n)^{-1}).$$

Consequently we have

$$\mathbb{P}\big(A_i(v) \cap B_i(u|v) \cap A_j(v') \cap B_j(u|v') \,|\, \mathcal{G}_n\big)$$
$$\leq \big(1 + O_p(\tilde{s}_n)\big)\frac{l^2 W(\lambda_u,\cdot)^{2\alpha} W(\lambda_v,\cdot)}{n^4\rho_n \mathcal{E}_W(\alpha)^2 \mathcal{E}_W}(1-a_{uv})(1-a_{uv'})$$

and therefore

$$\sum_{i=1}^{k}\sum_{j>i}\sum_{v,v'\in[n]\setminus\{u\}} \mathbb{P}\big(A_j(v') \cap B_j(u|v') \cap A_i(v) \cap B_i(u|v) \,|\, \mathcal{G}_n\big)$$
$$\leq (1 + O_p(\tilde{s}_n)) \cdot \sum_{i=1}^{k}\sum_{v\in[n]\setminus\{u\}} \frac{kl^2 W(\lambda_u,\cdot)^{2\alpha} W(\lambda_v,\cdot)}{n^3\rho_n \mathcal{E}_W(\alpha)^2 \mathcal{E}_W}(1-a_{uv})$$
$$\leq (1 + O_p(\tilde{s}_n)) \cdot \frac{klW(\lambda_u,\cdot)^{\alpha}}{n\rho_n \mathcal{E}_W(\alpha)} \cdot \frac{klW(\lambda_u,\cdot)^{\alpha}}{n\mathcal{E}_W(\alpha)\mathcal{E}_W} \cdot \frac{1}{n}\sum_{v\in[n]\setminus\{u\}} W(\lambda_v,\cdot)(1-a_{uv})$$
$$= (1 + O_p(\tilde{s}_n)) \cdot \frac{klW(\lambda_u,\cdot)^{\alpha}}{n\rho_n \mathcal{E}_W(\alpha)} \cdot \frac{klW(\lambda_u,\cdot)}{n\mathcal{E}_W(\alpha)\mathcal{E}_W} \cdot \int_0^1 (1-\rho_n W(\lambda_u,y))W(y,\cdot)\,dy,$$

where in the last line we have used Lemma 11. We then note that as we have

$$\sum_{i=1}^{k+1}\sum_{v\in[n]\setminus\{u\}} \mathbb{P}\big(A_i(v) \cap B_i(u|v) \,|\, \mathcal{G}_n\big)$$
$$= \big(1 + O_p\big(\tilde{s}_n(\gamma)\big)\big) \cdot \frac{(k+1)lW(\lambda_u,\cdot)^{\alpha}}{n\mathcal{E}_W(\alpha)\mathcal{E}_W} \cdot \int_0^1 (1-\rho_n W(\lambda_u,y))W(y,\cdot)\,dy,$$

by the formulae stated above and Lemma 11, we can therefore conclude by combining the above bounds that

$$\mathbb{P}\big(u \in \mathcal{V}(S_{ns}(\mathcal{G}_n) \setminus S_0(\mathcal{G}_n)) \,|\, \mathcal{G}_n\big)$$
$$= \big(1 + O_p(\tilde{s}_n)\big) \cdot \frac{(k+1)lW(\lambda_u,\cdot)^{\alpha}}{\mathcal{E}_W(\alpha)\mathcal{E}_W} \int_0^1 (1-\rho_n W(\lambda_u,y))W(y,\cdot)\,dy.$$

Consequently $\mathbb{P}\big(u \in S(\mathcal{G}_n) \,|\, \mathcal{G}_n\big)$ equals

$$\big(1 + O_p\big(\tilde{s}_n(\gamma)\big)\big) \cdot \frac{1}{n}\Big\{\frac{kW(\lambda_u,\cdot)}{\mathcal{E}_W} + \frac{(k+1)lW(\lambda_u,\cdot)^{\alpha}}{\mathcal{E}_W(\alpha)\mathcal{E}_W}\int_0^1 (1-\rho_n W(\lambda_u,y))W(y,\cdot)\,dy\Big\}$$

as desired. □

Proof of Theorem 6. We begin by highlighting that for the given model, we have that

$$W(\lambda,\cdot) = \frac{1}{k}\big(p + (k-1)q\big) = \mathcal{E}_W, \qquad W(\lambda,\cdot)^{\alpha} = \mathcal{E}_W(\alpha) \tag{59}$$

for $\lambda \in [0,1]$, and therefore we have that

$$f_n(\lambda_i,\lambda_j,1) = \frac{2k}{\kappa^{-1}(p+(\kappa-1)q)} = 2kc_1,$$
$$f_n(\lambda_i,\lambda_j,0) = 2l(k+1),$$
$$\tilde{g}_n(\lambda_i) = k + l(k+1)\cdot\big(1-\rho_n\kappa^{-1}(p+(\kappa-1)q)\big) = c_2.$$

Table 2: Summary statistics of Cora, CiteSeer and PubMedDiabetes datasets.

Dataset	Nodes	Edges	Features	Classes
Cora	2708	5429	1433	7
CiteSeer	3312	4732	3703	6
PubMed	19717	44338	500	3

Consequently, as a result of Proposition 2 viii), we know that we can obtain the minimizing kernel K_n^* which appears in the convergence theorem Theorem 2 as follows: we obtain a matrix $\tilde{K} \in \mathbb{R}^{k \times k}$ obtained via minimizing the convex function

$$-\frac{1}{\kappa^2}\sum_{i,j}\left\{2kc_1 \cdot (p\delta_{ij} + q(1-\delta_{ij}))\log\sigma(\tilde{K}_{ij}) + 2l(k+1)\log\sigma(-\tilde{K}_{ij})\right\} + \xi c_2 \sum_{i=1}^{\kappa}\tilde{K}_{ii}$$
$$= -\frac{1}{\kappa^2}\sum_{i,j}\left\{2kc_1 \cdot (p\delta_{ij} + q(1-\delta_{ij}))\log\sigma(\tilde{K}_{ij}) + 2l(k+1)\log\sigma(-\tilde{K}_{ij})\right\} + \xi c_2 \|\tilde{K}\|_*$$

over all positive semi-definite matrices $\tilde{K}$. The desired convergence then follows by applying Theorem 2. □

G Additional experimental details

We now describe the hyperparameter and training details of each of the methods used in the experiments; for all the methods, we used the Stellargraph[3] implementation of the architecture [20]. The code used to run the experiments is available on GitHub[4]. Experiments were run on a cluster, using for each experiment 4 cores of a Intel Xeon Gold 6126 2.6 Ghz CPU, and a variable amount of RAM depending on the method and dataset used. In total, the experiments carried out used approximately 100k CPU hours in total (including preliminary experiments). Table 2 gives a summary of the features of the Cora, CiteSeer and PubMedDiabetes datasets used in the experiments.

node2vec - We train node2vec with $p = q = 1$ for 5 epochs, using 50 random walks of length 5 per node to form as subsamples, and train using batch sizes of 64. We use an Adam algorithm with learning rate 10^{-3}. For the regularization, we use tf.regularizers.l2 with the specified regularization weight as the embeddings regularizer argument to the Embeddings layer used in the node2vec implementation.

Unlike as reported in [26], we found that using the Adam algorithm with learning rate 10^{-3} lead to far better performance than stochastic gradient descent with rates of either this magnitude or those suggested in e.g. the experiments performed within the GraphSAGE paper (of 0.2, 0.4 or 0.8). In our preliminary experiments, we generally found that varying the learning rates of the Adam method rarely lead to significant changes in performance and kept any observed trends relatively stable, and so we did not vary these significantly throughout.

GraphSAGE - For GraphSAGE, we used a two layer mean-pooling rule with neighbourhood sampling sizes of 10 and 5 respectively; we note that using 25 and 10 samples as suggested in [26] were computationally prohibitive for all the experiments we wished to carry out. Otherwise, we use the node2vec loss with 10 random walks of length 5 per node, use a batch size of 256 for training, and train for 10 epochs.

GCN - To train a GCN in an unsupervised fashion, we parameterize the embeddings in the usual node2vec loss through a two layer GCN with ReLU activations, with intermediate layer sizes 256 and 256. For the node2vec loss, we instead use 10 random walks of length 5 per node, use a batch size of 256 during training, and train the loss for 10 epochs.

DGI - For DGI, we use the same parameters as specified in [64]. We use 256 dimensional embeddings only; we found in our preliminary experiments the performance change in using a 512 dimensional embedding was negligible, and that on the PubMedDiabetes dataset, the memory usage required to

[3]We highlight that the Stellargraph package is licsensed under the Apache License 2.0.

[4]`https://github.com/aday651/embed-reg`

learn a 256 dimensional embedding was substantial (above 32GB of RAM needed). Otherwise, we train a one dimensional GCN with ReLU activation using the DGI methodology, for 100 epochs with an early stopping rule with a tolerance of 20 epochs, a batch size of 256, and used Adam with learning rate 10^{-3}.

Classifier details - Given the embeddings learned in an unsupervised fashion, there is then the need to build a classifier for both the node classification and link prediction tasks. To do so, we use logistic regression, namely the LogisticRegressionCV method from the scikit-learn Python package. The cross validation was set to use 5-folds and a 'one vs rest' classification scheme. Otherwise we used the default settings, except for a larger tolerance of the number of iterations for the BFGS optimization scheme.