Supplementary Material:

Graph Embedding with Shifted Inner Product Similarity and Its Improved Approximation Capability

A Experimental details

Visualization of Fig. 1: In Section 6.2, word feature vectors are computed from WordNet dataset. We used feature vectors computed by SIPS with $K = 5$. Since $(\boldsymbol{y}_i, u_i) \in \mathbb{R}^5$ for SIPS, we actually used $\boldsymbol{y}_i \in \mathbb{R}^4$ for the visualization. We extracted 97 words from the $n = 4027$ nouns, and applied t-SNE to $\{\boldsymbol{y}_i\}$ for the extracted words. Words with any hypernymy relations are connected by segments. In other words, v_i and v_j are connected when $w_{ij} = 1$. For extracting the 97 words, we chose the word "animal" as the root. Then chose four subordinate words ("mammal", "fish", "reptile", "invertebrate") connected to the root, and sampled more subordinate words from these four words, so that the total number of words becomes 97. Words are grouped by the four subordinate words of the root, which are indicated by the colors.

Optimization: In Section 6.1, all parameters are initialized as He et al. (2015) and trained by Adam (Kingma and Ba, 2014) with initial learning rate 0.01 and batch size 64. The number of iterations is 300,000. To ensure robust comparison, we save model parameters at every 5,000 iterations, and select the best performance parameters tested on the validation set. In Section 6.2, the most settings are the same as Section 6.1. All parameters are initialized as He et al. (2015) and trained by Adam with initial learning rate 0.001 and batch size 128. The number of iterations is 150,000.

B Relationship between the Poisson model and the Bernoulli model

For a pair $(i, j) \in \mathcal{I}_n$, we consider the Poisson model $w_{ij} \sim \mathrm{Po}(\lambda_{ij})$ with $\lambda_{ij} = \exp(h(\boldsymbol{x}_i, \boldsymbol{x}_j))$. In the below, w_{ij} and λ_{ij} are denoted as w and λ for simplifying the notation. Noting $P(w = k) = \exp(-\lambda)\lambda^k/k!$ for $k \in \{0, 1, \ldots, \}$, by Taylor expansion around $\lambda = 0$, we have $P(w = 0) = e^{-\lambda} = 1 - \lambda + \lambda^2/2 + O(\lambda^3)$ and $P(w = 1) = e^{-\lambda}\lambda = (1 - \lambda + O(\lambda^2))\lambda = \lambda - \lambda^2 + O(\lambda^3)$, and thus $P(w \geq 2) = 1 - P(w = 0) - P(w = 1) = \lambda^2/2 = O(\lambda^2)$. On the other hand, $\sigma(h(\boldsymbol{x}_i, \boldsymbol{x}_j)) = (1 + \lambda^{-1})^{-1} = \lambda - \lambda^2 + O(\lambda^3)$. Therefore, $P(w = 1) = \sigma(h(\boldsymbol{x}_i, \boldsymbol{x}_j)) + O(\lambda^3)$, proving (1).

When link weights are very sparse as is often seen in applications, most of λ_{ij}'s will be very small. Then the above results imply that $P(w_{ij} \geq 2) \approx 0$ can be ignored and $P(w_{ij} = 1) \approx \sigma(h(\boldsymbol{x}_i, \boldsymbol{x}_j))$ is interpreted as the Bernoulli model.

Let us consider a transformation from w_{ij} to $\tilde{w}_{ij} \in \{0, 1\}$ as $\tilde{w}_{ij} := \mathbf{1}(w_{ij} > 0)$. By noting $P(\tilde{w}_{ij} = 1) = P(w_{ij} > 0) = 1 - P(w_{ij} = 0) = \lambda_{ij} - \lambda_{ij}/2 + O(\lambda_{ij}^3)$, we have

$$P(\tilde{w}_{ij} = 1 \mid \boldsymbol{x}_i, \boldsymbol{x}_j) = \sigma(h(\boldsymbol{x}_i, \boldsymbol{x}_j)) + O(\lambda_{ij}^2).$$

Thus the Poisson model for w_{ij} is also interpreted as the Bernoulli model for the truncated variable $\tilde{w}_{ij}$.

C Proofs

C.1 Proof of Proposition 4.1

With $v = (2M)^{2p}$ and $\int = \int_{[-M,M]^p}$, a lower-bound of $\frac{1}{v}\iint |-\|\boldsymbol{x} - \boldsymbol{x}'\|_2^2 - \langle \boldsymbol{f}(\boldsymbol{x}), \boldsymbol{f}(\boldsymbol{x}')\rangle| \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{x}'$ is derived as

$$\begin{aligned}
\frac{1}{v}\iint \left| -\|\boldsymbol{x} - \boldsymbol{x}'\|_2^2 - \langle \boldsymbol{f}(\boldsymbol{x}), \boldsymbol{f}(\boldsymbol{x}')\rangle \right| \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{x}' &\geq \left| \frac{1}{v}\iint \left(-\|\boldsymbol{x} - \boldsymbol{x}'\|_2^2 - \langle \boldsymbol{f}(\boldsymbol{x}), \boldsymbol{f}(\boldsymbol{x}')\rangle\right) \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{x}' \right| \\
&= \left| \frac{1}{v}\iint \left(2\langle \boldsymbol{x}, \boldsymbol{x}'\rangle - \|\boldsymbol{x}\|_2^2 - \|\boldsymbol{x}'\|_2^2 - \langle \boldsymbol{f}(\boldsymbol{x}), \boldsymbol{f}(\boldsymbol{x}')\rangle\right) \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{x}' \right| \\
&= \left| \frac{1}{v}\left(2\left\|\int \boldsymbol{x}\mathrm{d}\boldsymbol{x}\right\|_2^2 - 2\int \mathrm{d}\boldsymbol{x}\int \|\boldsymbol{x}\|_2^2 \mathrm{d}\boldsymbol{x} - \left\|\int \boldsymbol{f}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right\|_2^2\right) \right|.
\end{aligned}$$

The terms in the last formula are computed as $\int \boldsymbol{x} \mathrm{d}\boldsymbol{x} = \boldsymbol{0}, \int \mathrm{d}\boldsymbol{x} = (2M)^p$,

$$\int \|\boldsymbol{x}\|_2^2 \mathrm{d}\boldsymbol{x} = \sum_{i=1}^{p} \int x_i^2 \mathrm{d}\boldsymbol{x} = (2M)^{p-1} \sum_{i=1}^{p} \int_{-M}^{M} x_i^2 \mathrm{d}x_i = (2M)^{p-1} \frac{2pM^3}{3} = (2M)^p \frac{pM^2}{3}.$$

Considering $\| \int f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \|_2^2 \geq 0$, we have

$$\frac{1}{v} \iint \Big| - \|\boldsymbol{x} - \boldsymbol{x}'\|_2^2 - \langle \boldsymbol{f}(\boldsymbol{x}), \boldsymbol{f}(\boldsymbol{x}') \rangle \Big| \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{x}' \geq \frac{2}{v} \int \mathrm{d}\boldsymbol{x} \int \|\boldsymbol{x}\|_2^2 \mathrm{d}\boldsymbol{x} = \frac{2pM^2}{3}.$$

Taking $\inf_{\boldsymbol{f} \in \mathfrak{S}(K)}$ proves the assertion.

□

C.2 Proof of Theorem 4.1 (Approximation theorem for SIPS)

Since $g_*^{(\mathrm{CPD})} : \mathcal{Y}^2 \to \mathbb{R}$ is a conditionally positive definite kernel on a compact set, Lemma 2.1 of Berg et al. (1984) indicates that

$$g_0(\boldsymbol{y}_*, \boldsymbol{y}_*') := g_*^{(\mathrm{CPD})}(\boldsymbol{y}_*, \boldsymbol{y}_*') - g_*^{(\mathrm{CPD})}(\boldsymbol{y}_*, \boldsymbol{y}_0) - g_*^{(\mathrm{CPD})}(\boldsymbol{y}_0, \boldsymbol{y}_*') + g_*^{(\mathrm{CPD})}(\boldsymbol{y}_0, \boldsymbol{y}_0)$$

is positive definite for arbitrary $\boldsymbol{y}_0 \in \mathcal{Y}$. We fix $\boldsymbol{y}_0$ in the argument below. According to Okuno et al. (2018) Theorem 5.1 (Theorem 3.2 in this paper), we can specify a neural network $\boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x})$ such that

$$\sup_{\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}} \Big| g_0\left(\boldsymbol{f}_*(\boldsymbol{x}), \boldsymbol{f}_*(\boldsymbol{x}')\right) - \langle \boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x}), \boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x}') \rangle \Big| < \varepsilon_1$$

for any ε_1. Next, let us consider a continuous function $h_*(\boldsymbol{x}) := g_*(\boldsymbol{f}_*(\boldsymbol{x}), \boldsymbol{y}_0) - \frac{1}{2} g_*(\boldsymbol{y}_0, \boldsymbol{y}_0)$. It follows from the universal approximation theorem (Cybenko, 1989; Telgarsky, 2017) that for any $\varepsilon_2 > 0$, there exists $m_u \in \mathbb{N}$ such that

$$\sup_{\boldsymbol{x} \in \mathcal{X}} |h_*(\boldsymbol{x}) - u_{\mathrm{NN}}(\boldsymbol{x})| < \varepsilon_2.$$

Therefore, we have

$$\begin{aligned}
&\sup_{\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}} \Big| g_*^{(\mathrm{CPD})}\left(\boldsymbol{f}_*(\boldsymbol{x}), \boldsymbol{f}_*(\boldsymbol{x}')\right) - \{\langle \boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x}), \boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x}') \rangle + u_{\mathrm{NN}}(\boldsymbol{x}) + u_{\mathrm{NN}}(\boldsymbol{x}')\} \Big| \\
&\quad = \sup_{\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}} \Big| \left(g_0\left(\boldsymbol{f}_*(\boldsymbol{x}), \boldsymbol{f}_*(\boldsymbol{x}')\right) - \langle \boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x}), \boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x}') \rangle\right) \\
&\qquad\qquad + (h_*(\boldsymbol{x}) - u_{\mathrm{NN}}(\boldsymbol{x})) + (h_*(\boldsymbol{x}') - u_{\mathrm{NN}}(\boldsymbol{x}')) \Big| \\
&\quad \leq \sup_{\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}} \Big| \left(g_0\left(\boldsymbol{f}_*(\boldsymbol{x}), \boldsymbol{f}_*(\boldsymbol{x}')\right) - \langle \boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x}), \boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x}') \rangle\right) \Big| \\
&\qquad + \sup_{\boldsymbol{x} \in \mathcal{X}} \Big| h_*(\boldsymbol{x}) - u_{\mathrm{NN}}(\boldsymbol{x}) \Big| + \sup_{\boldsymbol{x}' \in \mathcal{X}} \Big| h_*(\boldsymbol{x}') - u_{\mathrm{NN}}(\boldsymbol{x}') \Big| \\
&\quad < \varepsilon_1 + 2\varepsilon_2.
\end{aligned} \tag{17}$$

By letting $\varepsilon_1 = \varepsilon/2, \varepsilon_2 = \varepsilon/4$, the last formula becomes smaller than ε, thus proving

$$\sup_{\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}} \Big| g_*^{(\mathrm{CPD})}\left(\boldsymbol{f}_*(\boldsymbol{x}), \boldsymbol{f}_*(\boldsymbol{x}')\right) - \{\langle \boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x}), \boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x}') \rangle + u_{\mathrm{NN}}(\boldsymbol{x}) + u_{\mathrm{NN}}(\boldsymbol{x}')\} \Big| < \varepsilon.$$

□

C.3 Proof of Theorem 4.2 (Approximation theorem for C-SIPS)

With fixed $\boldsymbol{y}_0 \in \mathcal{Y}$, it follows from Berg et al. (1984) Lemma 2.1 and CPD-ness of the kernel $g_*^{(\mathrm{CPD})}$ that

$$g_0(\boldsymbol{y},\boldsymbol{y}') := g_*^{(\mathrm{CPD})}(\boldsymbol{y},\boldsymbol{y}') - g_*^{(\mathrm{CPD})}(\boldsymbol{y},\boldsymbol{y}_0) - g_*^{(\mathrm{CPD})}(\boldsymbol{y}_0,\boldsymbol{y}') + g_*^{(\mathrm{CPD})}(\boldsymbol{y}_0,\boldsymbol{y}_0)$$

is PD. Since $\mathcal{Y}$ is compact, we have $\sup_{\boldsymbol{y}\in\mathcal{Y}} |g_*^{(\mathrm{CPD})}(\boldsymbol{y},\boldsymbol{y}_0)| = a^2$ is bounded. Let us take a sufficiently large $r > a$ and define $\tau(\boldsymbol{y}) := \sqrt{r^2 + g_*^{(\mathrm{CPD})}(\boldsymbol{y},\boldsymbol{y}_0)}$. We consider a new kernel

$$g_1(\boldsymbol{y},\boldsymbol{y}') := g_0(\boldsymbol{y},\boldsymbol{y}') + 2\tau(\boldsymbol{y})\tau(\boldsymbol{y}').$$

Since both $g_0(\boldsymbol{y},\boldsymbol{y}')$ and $\tau(\boldsymbol{y})\tau(\boldsymbol{y}')$ are PD, $g_1(\boldsymbol{y},\boldsymbol{y}')$ is also PD. Applying Taylor's expansion $\sqrt{1+x} = 1 + x/2 + O(x^2)$, we have

$$\begin{aligned}
\tau(\boldsymbol{y})\tau(\boldsymbol{y}') &= \sqrt{r^2 + g_*^{(\mathrm{CPD})}(\boldsymbol{y},\boldsymbol{y}_0)}\sqrt{r^2 + g_*^{(\mathrm{CPD})}(\boldsymbol{y}',\boldsymbol{y}_0)} \\
&= r^2\sqrt{1 + g_*^{(\mathrm{CPD})}(\boldsymbol{y},\boldsymbol{y}_0)/r^2}\sqrt{1 + g_*^{(\mathrm{CPD})}(\boldsymbol{y}',\boldsymbol{y}_0)/r^2} \\
&= r^2(1 + g_*^{(\mathrm{CPD})}(\boldsymbol{y},\boldsymbol{y}_0)/2r^2 + O(r^{-4}))(1 + g_*^{(\mathrm{CPD})}(\boldsymbol{y}',\boldsymbol{y}_0)/2r^2 + O(r^{-4})) \\
&= r^2 + \frac{1}{2}(g_*^{(\mathrm{CPD})}(\boldsymbol{y},\boldsymbol{y}_0) + g_*^{(\mathrm{CPD})}(\boldsymbol{y}',\boldsymbol{y}_0)) + O(r^{-2}),
\end{aligned}$$

thus proving

$$g_1(\boldsymbol{y},\boldsymbol{y}') = g_0(\boldsymbol{y},\boldsymbol{y}') + 2\tau(\boldsymbol{y})\tau(\boldsymbol{y}') = g_*^{(\mathrm{CPD})}(\boldsymbol{y},\boldsymbol{y}') + g_*^{(\mathrm{CPD})}(\boldsymbol{y}_0,\boldsymbol{y}_0) + 2r^2 + O(r^{-2}).$$

Let us define $\gamma := g_*^{(\mathrm{CPD})}(\boldsymbol{y}_0,\boldsymbol{y}_0) + 2r^2 = O(r^2)$. Considering the PD-ness of $g_1(\boldsymbol{y},\boldsymbol{y}') = g_*^{(\mathrm{CPD})}(\boldsymbol{y},\boldsymbol{y}') + \gamma + O(r^{-2})$, we have

$$\begin{aligned}
&\sup_{\boldsymbol{x},\boldsymbol{x}'\in\mathcal{X}} \left| g_*^{(\mathrm{CPD})}(\boldsymbol{f}_*(\boldsymbol{x}),\boldsymbol{f}_*(\boldsymbol{x}')) - (\langle \boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x}), \boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x}')\rangle - \gamma) \right| \\
&= \sup_{\boldsymbol{x},\boldsymbol{x}'\in\mathcal{X}} \left| g_1(\boldsymbol{f}_*(\boldsymbol{x}),\boldsymbol{f}_*(\boldsymbol{x}')) - \langle \boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x}), \boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x}')\rangle \right| + O(r^{-2}) \qquad (18) \\
&< \varepsilon + O(r^{-2}).
\end{aligned}$$

□

D Approximation Error Rate

We first discuss the approximation error rate for truncating the series expansion of Mercer's theorem in Section D.1 and the approximation error rate for NNs in Section D.2. Then, by considering these error rates, we prove Theorems 5.1 and 5.2 for IPS and SIPS, respectively, in Sections D.3 and D.4.

D.1 Error rate for Mercer's theorem

We evaluate the error rate for Mercer's theorem (shown as Theorem 3.1 in this paper) to approximate PD kernels g_* satisfying conditions (C-1) and (C-2) of Section 5.

We define the error rate for Mercer's theorem as

$$\varepsilon_1(K) := \sup_{\boldsymbol{y},\boldsymbol{y}'\in\mathcal{Y}} \left| g_*(\boldsymbol{y},\boldsymbol{y}') - \sum_{k=1}^{K} \lambda_k \phi_k(\boldsymbol{y})\phi_k(\boldsymbol{y}') \right|. \qquad (19)$$

Then, the error rate is given in the lemma below.

Lemma D.1 For compact set $\mathcal{Y} \subset \mathbb{R}^{K^*}$, $K^* \in \mathbb{N}$, we consider a PD kernel $g_* : \mathcal{Y}^2 \to \mathbb{R}$ which satisfies conditions (C-1) and (C-2). Then, $\varepsilon_1(K) = O(K^{-1/K^*})$.

For proving the lemma, we first show a result of the decay rate for eigenvalues. The theorem below is a special case of Theorem 4 of Cobos and Kühn (1990) by assuming μ as Lebesgue measure, and $\boldsymbol{\Omega} = \mathcal{Y}$.

Theorem D.1 (Cobos and Kühn (1990)) Let $\mathcal{Y} \subset \mathbb{R}^L$ be a non-empty compact set for $L \in \mathbb{N}$, and let $g : \mathcal{Y}^2 \to \mathbb{R}$ be a positive definite kernel satisfying $\int_{\mathcal{Y}} \|g(\boldsymbol{t}, \cdot)\|_{C^\alpha} \mathrm{d}\boldsymbol{t} < \infty$, where $0 < \alpha \leq 1$ and

$$\|g(\boldsymbol{t}, \cdot)\|_{C^\alpha} := \max \left\{ \sup_{\boldsymbol{y} \in \mathcal{Y}} |g(\boldsymbol{t}, \boldsymbol{y})|, \sup_{\substack{\boldsymbol{y}, \boldsymbol{y}' \in \mathcal{Y} \\ \boldsymbol{y} \neq \boldsymbol{y}'}} \frac{|g(\boldsymbol{t}, \boldsymbol{y}) - g(\boldsymbol{t}, \boldsymbol{y}')|}{\|\boldsymbol{y} - \boldsymbol{y}'\|_2^\alpha} \right\}.$$

Then, the k-th largest eigenvalue of g is

$$\lambda_k = O(k^{-1-\alpha/L}).$$

We apply Theorem D.1 to g_* by letting $L = K^*$ and $\alpha = 1$. Then the eigenvalues of g_* satisfy

$$\lambda_k = O(k^{-1-1/K^*}), \tag{20}$$

where the condition of g in Theorem D.1 will be verified later. On the other hand, Mercer's theorem and the condition (C-1) leads to

$$\begin{aligned} \varepsilon_1(K) = \sup_{\boldsymbol{y}, \boldsymbol{y}' \in \mathcal{Y}} \left| \sum_{k=K+1}^{\infty} \lambda_k \phi_k(\boldsymbol{y}) \phi_k(\boldsymbol{y}') \right| &\leq \sum_{k=K+1}^{\infty} \lambda_k \sup_{\boldsymbol{y} \in \mathcal{Y}, l \in \mathbb{N}} |\phi_l(\boldsymbol{y})| \sup_{\boldsymbol{y}' \in \mathcal{Y}, l' \in \mathbb{N}} |\phi_{l'}(\boldsymbol{y}')| \\ &= \left(\sup_{\boldsymbol{y} \in \mathcal{Y}, k \in \mathbb{N}} |\phi_k(\boldsymbol{y})| \right)^2 \sum_{k=K+1}^{\infty} \lambda_k = O\left(\sum_{k=K+1}^{\infty} \lambda_k \right). \end{aligned} \tag{21}$$

Therefore, substituting (20) into (21), we have

$$\varepsilon_1(K) = O\left(\sum_{k=K+1}^{\infty} \lambda_k \right) = O\left(\int_K^\infty k^{-1-1/K^*} \mathrm{d}k \right) = O\left(\left[-K^* k^{-1/K^*} \right]_K^\infty \right) = O(K^{-1/K^*}).$$

This proves Lemma D.1. Finally, we verify that g_* satisfies the condition of g in Theorem D.1. As g_* is continuous on compact set,

$$\sup_{\boldsymbol{t} \in \mathcal{Y}} \sup_{\boldsymbol{y} \in \mathcal{Y}} |g_*(\boldsymbol{t}, \boldsymbol{y})| < \infty \tag{22}$$

obviously holds, and the condition (C-2) implies α-Hölder continuity, and so

$$\sup_{\boldsymbol{t} \in \mathcal{Y}} \sup_{\substack{\boldsymbol{y}, \boldsymbol{y}' \in \mathcal{Y} \\ \boldsymbol{y} \neq \boldsymbol{y}'}} \frac{|g_*(\boldsymbol{t}, \boldsymbol{y}) - g_*(\boldsymbol{t}, \boldsymbol{y}')|}{\|\boldsymbol{y} - \boldsymbol{y}'\|_2} < \infty. \tag{23}$$

Inequalities (22) and (23) lead to

$$\sup_{\boldsymbol{t} \in \mathcal{Y}} \|g_*(\boldsymbol{t}, \cdot)\|_{C^1} \leq \max \left\{ \sup_{\boldsymbol{t} \in \mathcal{Y}} \sup_{\boldsymbol{y} \in \mathcal{Y}} |g_*(\boldsymbol{t}, \boldsymbol{y})|, \sup_{\boldsymbol{t} \in \mathcal{Y}} \sup_{\substack{\boldsymbol{y}, \boldsymbol{y}' \in \mathcal{Y} \\ \boldsymbol{y} \neq \boldsymbol{y}'}} \frac{|g_*(\boldsymbol{t}, \boldsymbol{y}) - g_*(\boldsymbol{t}, \boldsymbol{y}')|}{\|\boldsymbol{y} - \boldsymbol{y}'\|_2} \right\} < \infty.$$

Thus g_* satisfies

$$\int_{\mathcal{Y}} \|g_*(\boldsymbol{t}, \cdot)\|_{C^1} \mathrm{d}\boldsymbol{t} \leq \sup_{\boldsymbol{t} \in \mathcal{Y}} \|g_*(\boldsymbol{t}, \cdot)\|_{C^1} \int_{\mathcal{Y}} \mathrm{d}\boldsymbol{t} < \infty,$$

because compact set $\mathcal{Y} \subset \mathbb{R}^{K^*}$ is bounded and closed. □

D.2 Error rate for NN approximations

We refer to the result of Yarotsky (2018). By combining Proposition 1 ($\alpha = 0$, i.e., constant-depth shallow NNs) and Theorem 2 ($0 < \alpha \le 1$, i.e., deep NNs with growing depth as W increases) of Yarotsky (2018), we have the following theorem.

Theorem D.2 (Yarotsky (2018)) For $\mathcal{X} = [-M, M]^p$, $M > 0$, $p \in \mathbb{N}$ and $0 \le \alpha \le 1$, we consider the set of real-valued NNs $v_{\mathrm{NN}} \in \mathfrak{S}_\alpha(W, 1)$ for $W \in \mathbb{N}$. Let $\omega(v; r) := \max\{|v(\boldsymbol{x}) - v(\boldsymbol{x}')| : \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}, \|\boldsymbol{x} - \boldsymbol{x}'\| \le r\}$ be the modulus of continuity. Then, there exist $a, c \in \mathbb{R}$ such that

$$\inf_{v_{\mathrm{NN}} \in \mathfrak{S}_\alpha(W,1)} \sup_{\boldsymbol{x} \in \mathcal{X}} |v_*(\boldsymbol{x}) - v_{\mathrm{NN}}(\boldsymbol{x})| \le a\omega(v_*; cW^{-\frac{1+\alpha}{p}})$$

holds for any real-valued continuous function $v_* : \mathcal{X} \to \mathbb{R}$.

In later sections, we will use the following two lemmas, which are immediate consequences of Theorem D.2.

Lemma D.2 Symbols are the same as those of Theorem D.2. Assume that v_* is continuously differentiable over $\mathcal{X}$, and fix such a v^*. Then, as $W \to \infty$, we have

$$\inf_{v_{\mathrm{NN}} \in \mathfrak{S}_\alpha(W,1)} \sup_{\boldsymbol{x} \in \mathcal{X}} |v_*(\boldsymbol{x}) - v_{\mathrm{NN}}(\boldsymbol{x})| = O(W^{-\frac{1+\alpha}{p}}).$$

Proof is based on the intermediate value theorem. For $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$ satisfying $\|\boldsymbol{x} - \boldsymbol{x}'\| \le r$, there exists $\boldsymbol{x}_0 \in \mathcal{X}$ such that $v_*(\boldsymbol{x}) - v_*(\boldsymbol{x}') = \frac{\partial v_*(\boldsymbol{x})}{\partial \boldsymbol{x}}|_{\boldsymbol{x}=\boldsymbol{x}_0}(\boldsymbol{x} - \boldsymbol{x}')$. Since $b := \sup_{\boldsymbol{x} \in \mathcal{X}} \|\partial v_*(\boldsymbol{x})/\partial \boldsymbol{x}\|$ is bounded because of the continuity of the first-order derivative $\partial v_*(\boldsymbol{x})/\partial \boldsymbol{x}$, Cauchy-Schwarz inequality indicates

$$|v_*(\boldsymbol{x}) - v_*(\boldsymbol{x}')| \le \left\| \frac{\partial v_*(\boldsymbol{x})}{\partial \boldsymbol{x}} \bigg|_{\boldsymbol{x}=\boldsymbol{x}_0} \right\|_2 \|\boldsymbol{x} - \boldsymbol{x}'\|_2 \le br.$$

Thus we have $\omega(v_*; r) \le br$, indicating

$$a\omega(v_*; cW^{-\frac{1+\alpha}{p}}) \le abcW^{-\frac{1+\alpha}{p}}. \tag{24}$$

Substituting (24) into Theorem D.2 proves the lemma. □

Lemma D.3 For $\mathcal{X} = [-M, M]^p$, $M > 0$, $p \in \mathbb{N}$ and $0 \le \alpha \le 1$, we consider the set of NNs $\boldsymbol{v}_{\mathrm{NN}} \in \mathfrak{S}_\alpha(W, K)$ for $W, K \in \mathbb{N}$. Let $\boldsymbol{v}_* : \mathcal{X} \to \mathbb{R}^K$ be a vector-valued continuously differentiable function over $\mathcal{X}$ such that $\sup_{k \in \{1,\dots,K\}, \boldsymbol{x} \in \mathcal{X}} \|\partial v_{*k}(\boldsymbol{x})/\partial \boldsymbol{x}\|_2 \le b$ for some b which does not depend on K. Then, as $W/K \to \infty$, we have

$$\inf_{\boldsymbol{v}_{\mathrm{NN}} \in \mathfrak{S}_\alpha(W,K)} \sup_{\boldsymbol{x} \in \mathcal{X}} \|\boldsymbol{v}_*(\boldsymbol{x}) - \boldsymbol{v}_{\mathrm{NN}}(\boldsymbol{x})\|_2 = O(K^{\frac{1}{2}+\frac{1+\alpha}{p}} W^{-\frac{1+\alpha}{p}}).$$

Proof is based on applying Lemma D.2 to each of K output units of $\boldsymbol{v}_*$. We consider K real-valued neural networks of depth $L = O((W/K)^\alpha)$ with W/K weights as shown in Fig. 2. Since such NNs are included in $\mathfrak{S}_\alpha(W, K)$, we have

$$\inf_{\boldsymbol{v}_{\mathrm{NN}} \in \mathfrak{S}_\alpha(W,K)} \sup_{\boldsymbol{x} \in \mathcal{X}} \|\boldsymbol{v}_*(\boldsymbol{x})) - \boldsymbol{v}_{\mathrm{NN}}(\boldsymbol{x})\|_2 \le \left(\sum_{k=1}^K \inf_{v_k \in \mathfrak{S}_\alpha(W/K,1)} \sup_{\boldsymbol{x} \in \mathcal{X}} |v_{*k}(\boldsymbol{x}) - v_k(\boldsymbol{x})|^2 \right)^{1/2},$$

where $\boldsymbol{v}_*(\boldsymbol{x}) = (v_{*1}(\boldsymbol{x}), v_{*2}(\boldsymbol{x}), \dots, v_{*K}(\boldsymbol{x})), \boldsymbol{v}_{\mathrm{NN}}(\boldsymbol{x}) = (v_1(\boldsymbol{x}), v_2(\boldsymbol{x}), \dots, v_K(\boldsymbol{x}))$. We apply Lemma D.2 with W/K weights to each v_{*k}, where the same bound b is used in (24). Then the error is bounded by $\sqrt{K} \times abc(W/K)^{-\frac{1+\alpha}{p}} = O(K^{\frac{1}{2}+\frac{1+\alpha}{p}} W^{-\frac{1+\alpha}{p}})$. □

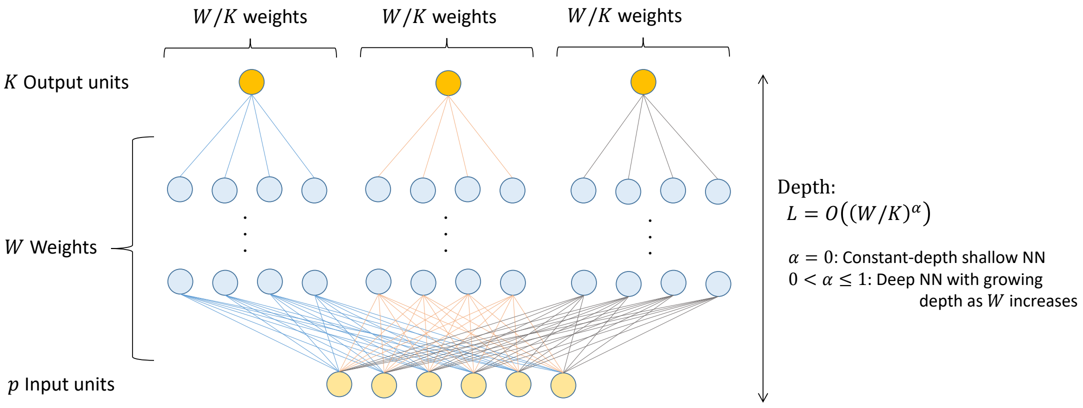
W/K weights
W/K weights
W/K weights
K Output units
W Weights
p Input units
Depth:
$L = O\big((W/K)^\alpha\big)$
$\alpha = 0$: Constant-depth shallow NN
$0 < \alpha \le 1$: Deep NN with growing depth as W increases

Figure 2: A structure of vector-valued neural network $\boldsymbol{v}_{\text{NN}} : \mathbb{R}^p \to \mathbb{R}^K$ having W weights. We allocate W/K weights to each output unit, so that weights are not shared by the K output units. In practice, internal units are often shared by the output units, but we consider the above structure for showing the upper bound of the approximation error.

D.3 Proof of Theorem 5.1 (Approximation error rate for IPS)

Applying Theorem 3.1 to a PD kernel $g_*^{(\text{PD})}$, there exist eigenvalues $\{\lambda_k\}_{k=1}^\infty$, $\lambda_1 \ge \lambda_2 \ge \cdots$ and eigenfunctions $\{\phi_k(\boldsymbol{y})\}_{k=1}^\infty$ such that $\sum_{k=1}^K \lambda_k \phi_k(\boldsymbol{y})\phi_k(\boldsymbol{y}')$ absolutely and uniformly converges to $g_*^{(\text{PD})}(\boldsymbol{y},\boldsymbol{y}')$ as $K \to \infty$. Here, we define two vector-valued functions

$$\begin{aligned}\boldsymbol{\eta}_K(\boldsymbol{y}) &:= (\lambda_1^{1/2}\phi_1(\boldsymbol{y}), \lambda_2^{1/2}\phi_2(\boldsymbol{y}), \dots, \lambda_K^{1/2}\phi_K(\boldsymbol{y})),\\ \tilde{\boldsymbol{\phi}}_K(\boldsymbol{x}) &:= \boldsymbol{\eta}_K(\boldsymbol{f}_*(\boldsymbol{x})),\end{aligned}$$

so that $\langle \boldsymbol{\eta}_K(\boldsymbol{f}_*(\boldsymbol{x})), \boldsymbol{\eta}_K(\boldsymbol{f}_*(\boldsymbol{x}'))\rangle = \langle \tilde{\boldsymbol{\phi}}_K(\boldsymbol{x}), \tilde{\boldsymbol{\phi}}_K(\boldsymbol{x}')\rangle = \sum_{k=1}^K \lambda_k \phi_k(\boldsymbol{f}_*(\boldsymbol{x}))\phi_k(\boldsymbol{f}_*(\boldsymbol{x}'))$. Using these functions, for any $\boldsymbol{f}_{\text{NN}} \in \mathfrak{S}_\alpha(W_f, K)$, we have

$$\begin{aligned}&\left|g_*^{(\text{PD})}(\boldsymbol{f}_*(\boldsymbol{x}), \boldsymbol{f}_*(\boldsymbol{x}')) - \langle \boldsymbol{f}_{\text{NN}}(\boldsymbol{x}), \boldsymbol{f}_{\text{NN}}(\boldsymbol{x}')\rangle\right|\\ &\quad\le \left|g_*(\boldsymbol{f}_*(\boldsymbol{x}), \boldsymbol{f}_*(\boldsymbol{x}')) - \langle \boldsymbol{\eta}_K(\boldsymbol{f}_*(\boldsymbol{x})), \boldsymbol{\eta}_K(\boldsymbol{f}_*(\boldsymbol{x}'))\rangle\right| \qquad (25)\\ &\qquad + \left|\langle \tilde{\boldsymbol{\phi}}_K(\boldsymbol{x}), \tilde{\boldsymbol{\phi}}_K(\boldsymbol{x}')\rangle - \langle \tilde{\boldsymbol{\phi}}_K(\boldsymbol{x}), \boldsymbol{f}_{\text{NN}}(\boldsymbol{x}')\rangle\right| + \left|\langle \tilde{\boldsymbol{\phi}}_K(\boldsymbol{x}), \boldsymbol{f}_{\text{NN}}(\boldsymbol{x}')\rangle - \langle \boldsymbol{f}_{\text{NN}}(\boldsymbol{x}), \boldsymbol{f}_{\text{NN}}(\boldsymbol{x}')\rangle\right|. \qquad (26)\end{aligned}$$

These terms (25) and (26) can be evaluated in the following way.

- Regarding the term (25),

$$\begin{aligned}&\sup_{\boldsymbol{x},\boldsymbol{x}'\in\mathcal{X}} \left|g_*^{(\text{PD})}(\boldsymbol{f}_*(\boldsymbol{x}), \boldsymbol{f}_*(\boldsymbol{x}')) - \langle \boldsymbol{\eta}_K(\boldsymbol{f}_*(\boldsymbol{x})), \boldsymbol{\eta}_K(\boldsymbol{f}_*(\boldsymbol{x}'))\rangle\right|\\ &\quad\le \sup_{\boldsymbol{y},\boldsymbol{y}'\in\mathcal{Y}} \left|g_*^{(\text{PD})}(\boldsymbol{y},\boldsymbol{y}') - \langle \boldsymbol{\eta}_K(\boldsymbol{y}), \boldsymbol{\eta}_K(\boldsymbol{y}')\rangle\right|\\ &\quad= \sup_{\boldsymbol{y},\boldsymbol{y}'\in\mathcal{Y}} \left|g_*^{(\text{PD})}(\boldsymbol{y},\boldsymbol{y}') - \sum_{k=1}^K \lambda_k\phi_k(\boldsymbol{y})\phi_k(\boldsymbol{y}')\right| = O(K^{-1/K^*}),\end{aligned}$$

where the last formula follows by applying Lemma D.1 to $g_*^{(\mathrm{PD})}$. Thus, we have

$$\inf_{\boldsymbol{f}_{\mathrm{NN}} \in \mathfrak{S}_\alpha(W_f, K)} \sup_{\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}} \left| g_*^{(\mathrm{PD})}(\boldsymbol{f}_*(\boldsymbol{x}), \boldsymbol{f}_*(\boldsymbol{x}')) - \langle \boldsymbol{\eta}_K(\boldsymbol{f}_*(\boldsymbol{x})), \boldsymbol{\eta}_K(\boldsymbol{f}_*(\boldsymbol{x}')) \rangle \right| = O(K^{-1/K^*}). \tag{27}$$

- Regarding the term (26),

$$\begin{aligned}
&\sup_{\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}} \left\{ \left| \langle \tilde{\boldsymbol{\phi}}_K(\boldsymbol{x}), \tilde{\boldsymbol{\phi}}_K(\boldsymbol{x}') \rangle - \langle \tilde{\boldsymbol{\phi}}_K(\boldsymbol{x}), \boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x}') \rangle \right| + \left| \langle \tilde{\boldsymbol{\phi}}_K(\boldsymbol{x}), \boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x}') \rangle - \langle \boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x}), \boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x}') \rangle \right| \right\} \\
&\leq \sup_{\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}} \left\{ \|\tilde{\boldsymbol{\phi}}_K(\boldsymbol{x})\|_2 \|\tilde{\boldsymbol{\phi}}_K(\boldsymbol{x}') - \boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x}')\|_2 + \|\boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x}')\|_2 \|\tilde{\boldsymbol{\phi}}_K(\boldsymbol{x}) - \boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x})\|_2 \right\} \\
&\leq \sup_{\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}} \left\{ \|\tilde{\boldsymbol{\phi}}_K(\boldsymbol{x})\|_2 \|\tilde{\boldsymbol{\phi}}_K(\boldsymbol{x}') - \boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x}')\|_2 + (\|\tilde{\boldsymbol{\phi}}_K(\boldsymbol{x}')\|_2 + \|\tilde{\boldsymbol{\phi}}_K(\boldsymbol{x}') - \boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x}')\|_2) \|\tilde{\boldsymbol{\phi}}_K(\boldsymbol{x}) - \boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x})\|_2 \right\} \\
&= 2 \sup_{\boldsymbol{x} \in \mathcal{X}} \|\tilde{\boldsymbol{\phi}}_K(\boldsymbol{x})\|_2 \sup_{\boldsymbol{x}' \in \mathcal{X}} \|\tilde{\boldsymbol{\phi}}_K(\boldsymbol{x}') - \boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x}')\|_2 + \sup_{\boldsymbol{x} \in \mathcal{X}} \|\tilde{\boldsymbol{\phi}}_K(\boldsymbol{x}) - \boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x})\|_2^2.
\end{aligned}$$

Here, $\|\tilde{\boldsymbol{\phi}}_K(\boldsymbol{x})\|_2 = \| \sum_{k=1}^K \lambda_k \phi_k(\boldsymbol{f}_*(\boldsymbol{x})) \phi_k(\boldsymbol{f}_*(\boldsymbol{x}))\|_2 \leq \| \sum_{k=1}^\infty \lambda_k \phi_k(\boldsymbol{f}_*(\boldsymbol{x})) \phi_k(\boldsymbol{f}_*(\boldsymbol{x}))\|_2 = \|g_*^{(\mathrm{PD})}(\boldsymbol{f}_*(\boldsymbol{x}), \boldsymbol{f}_*(\boldsymbol{x}))\|_2$ is bounded, because $g_*^{(\mathrm{PD})}(\boldsymbol{f}_*(\boldsymbol{x}), \boldsymbol{f}_*(\boldsymbol{x}))$ is continuous over the compact set $\mathcal{X}^2$. For applying Lemma D.3 to $\tilde{\boldsymbol{\phi}}_K(\boldsymbol{x})$, we need to show that the constant b exists. Noting $\|\partial \tilde{\phi}_k / \partial \boldsymbol{x}\|_2^2 = \sum_{i=1}^p (\partial \tilde{\phi}_k / \partial x_i)^2 \leq \sum_{i=1}^p \lambda_k \|\partial \phi_k / \partial \boldsymbol{y}\|_2^2 \, \|\partial \boldsymbol{f}_* / \partial x_i\|_2^2$, we have

$$\sup_{k \in \mathbb{N}} \sup_{\boldsymbol{x} \in \mathcal{X}} \|\partial \tilde{\phi}_k / \partial \boldsymbol{x}\|_2^2 \leq \sup_{k \in \mathbb{N}} \sup_{\boldsymbol{y} \in \mathcal{Y}} \lambda_k \|\partial \phi_k / \partial \boldsymbol{y}\|_2^2 \sup_{\boldsymbol{x} \in \mathcal{X}} \sum_{i=1}^p \|\partial \boldsymbol{f}_* / \partial x_i\|_2^2 < \infty, \tag{28}$$

where $\sup_{k \in \mathbb{N}} \sup_{\boldsymbol{y} \in \mathcal{Y}} \lambda_k \|\partial \phi_k / \partial \boldsymbol{y}\|_2^2 < \infty$ follows from (C-1) and $\sup_{\boldsymbol{x} \in \mathcal{X}} \sum_{i=1}^p \|\partial \boldsymbol{f}_* / \partial x_i\|_2^2 < \infty$ follows from (C-3). We can take b^2 as the upper bound of (28), and then Lemma D.3 implies

$$\inf_{\boldsymbol{f}_{\mathrm{NN}} \in \mathfrak{S}_\alpha(W_f, K)} \sup_{\boldsymbol{x} \in \mathcal{X}} \|\tilde{\boldsymbol{\phi}}_K(\boldsymbol{x}) - \boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x})\|_2 = O(K^{\frac{1}{2} + \frac{1+\alpha}{p}} W_f^{-\frac{1+\alpha}{p}}) \tag{29}$$

so that the evaluation of (26) leads to

$$\begin{aligned}
&\inf_{\boldsymbol{f}_{\mathrm{NN}} \in \mathfrak{S}_\alpha(W_f, K)} \sup_{\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}} \left\{ \left| \langle \tilde{\boldsymbol{\phi}}_K(\boldsymbol{x}), \tilde{\boldsymbol{\phi}}_K(\boldsymbol{x}') \rangle - \langle \tilde{\boldsymbol{\phi}}_K(\boldsymbol{x}), \boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x}') \rangle \right| + \left| \langle \tilde{\boldsymbol{\phi}}_K(\boldsymbol{x}), \boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x}') \rangle - \langle \boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x}), \boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x}') \rangle \right| \right\} \\
&= O(K^{\frac{1}{2} + \frac{1+\alpha}{p}} W_f^{-\frac{1+\alpha}{p}}).
\end{aligned} \tag{30}$$

Considering (27) and (30), we finally obtain

$$\inf_{\boldsymbol{f}_{\mathrm{NN}} \in \mathfrak{S}_\sigma(W_f, K)} \sup_{\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}} \left| g_*^{(\mathrm{PD})}(\boldsymbol{f}_*(\boldsymbol{x}), \boldsymbol{f}_*(\boldsymbol{x}')) - \langle \boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x}), \boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x}') \rangle \right| = O\Big(K^{-\frac{1}{K^*}} + K^{\frac{1}{2} + \frac{1+\alpha}{p}} W_f^{-\frac{1+\alpha}{p}} \Big).$$

□

D.4 Proof of Theorem 5.2 (Approximation error rate for SIPS)

Recall the inequality (17) in Section C.2.

$$\begin{aligned}
&\sup_{\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}} \left| g_*^{(\mathrm{CPD})}(\boldsymbol{f}_*(\boldsymbol{x}), \boldsymbol{f}_*(\boldsymbol{x}')) - (\langle \boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x}), \boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x}') \rangle + u_{\mathrm{NN}}(\boldsymbol{x}) + u_{\mathrm{NN}}(\boldsymbol{x}')) \right| \\
&\qquad \leq \sup_{\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}} \left| g_0(\boldsymbol{f}_*(\boldsymbol{x}), \boldsymbol{f}_*(\boldsymbol{x}')) - \langle \boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x}), \boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x}') \rangle \right| + 2 \sup_{\boldsymbol{x} \in \mathcal{X}} \left| h_*(\boldsymbol{x}) - u_{\mathrm{NN}}(\boldsymbol{x}) \right|
\end{aligned} \tag{31}$$

We evaluate the two terms in (31). Since we have assumed that $g_*^{(\mathrm{CPD})}$ is C^1 (the condition C-2), g_0 and h_* are also C^1. Then, by applying Theorem 5.1 to the PD kernel g_0, the first term in (31) is evaluated as

$$\inf_{\boldsymbol{f}_{\mathrm{NN}} \in \mathfrak{S}_\alpha(W_f, K)} \sup_{\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}} \left| g_0(\boldsymbol{f}_*(\boldsymbol{x}), \boldsymbol{f}_*(\boldsymbol{x}')) - \langle \boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x}), \boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x}') \rangle \right| = O\Big(K^{-\frac{1}{K^*}} + K^{\frac{1}{2} + \frac{1+\alpha}{p}} W_f^{-\frac{1+\alpha}{p}} \Big). \tag{32}$$

By applying Lemma D.2 to h_*, the second term in (31) is evaluated as

$$\inf_{u_{\mathrm{NN}}\in\mathfrak{S}_\alpha(W_u,1)} \sup_{\boldsymbol{x}\in\mathcal{X}} \left| h_*(\boldsymbol{x}) - u_{\mathrm{NN}}(\boldsymbol{x}) \right| = O\Big(W_u^{-\frac{1+\alpha}{p}}\Big). \tag{33}$$

Considering (31), (32) and (33), we obtain

$$\begin{aligned}
&\inf_{\substack{\boldsymbol{f}_{\mathrm{NN}}\in\mathfrak{S}_\alpha(W_f,K)\\ u_{\mathrm{NN}}\in\mathfrak{S}_\alpha(W_u,1)}} \sup_{\boldsymbol{x},\boldsymbol{x}'\in\mathcal{X}} \left| g_*^{(\mathrm{CPD})}(\boldsymbol{f}_*(\boldsymbol{x}),\boldsymbol{f}_*(\boldsymbol{x}')) - (\langle \boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x}),\boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x}')\rangle + u_{\mathrm{NN}}(\boldsymbol{x}) + u_{\mathrm{NN}}(\boldsymbol{x}')) \right| \\
&\quad = O\Big(K^{-\frac{1}{K^*}} + K^{\frac{1}{2}+\frac{1+\alpha}{p}} W_f^{-\frac{1+\alpha}{p}} + W_u^{-\frac{1+\alpha}{p}}\Big).
\end{aligned}$$

□

E Non-CPD Similarities

CPD includes a broad range of kernels, but there exists a variety of non-CPD kernels. One example is Epanechnikov kernel $g(\boldsymbol{y},\boldsymbol{y}') := (1-\|\boldsymbol{y}-\boldsymbol{y}'\|_2^2)\mathbf{1}(\|\boldsymbol{y}-\boldsymbol{y}'\|_2 \le 1)$. To approximate similarities based on such non-CPD kernels, we propose a novel model, yet based on inner product, with high approximation capability beyond SIPS. Although parameter optimization of this model is not always easy due to the excessive degrees of freedom, the model is, in theory, shown to be capable of approximating more general kernels that are considered in Ong et al. (2004).

E.1 Proposed model

Let us consider a similarity $h(\boldsymbol{x},\boldsymbol{x}') = g_*(f_*(\boldsymbol{x}), f_*(\boldsymbol{x}'))$ with any kernel $g_* : \mathbb{R}^{2K^*} \to \mathbb{R}$ and a continuous map $f_* : \mathbb{R}^p \to \mathbb{R}^{K^*}$. To approximate it, we consider a similarity model

$$h(\boldsymbol{x}_i,\boldsymbol{x}_j) = \langle \boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x}_i), \boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x}_j)\rangle - \langle \boldsymbol{r}_{\mathrm{NN}}(\boldsymbol{x}_i), \boldsymbol{r}_{\mathrm{NN}}(\boldsymbol{x}_j)\rangle, \tag{34}$$

where $\boldsymbol{f}_{\mathrm{NN}} : \mathbb{R}^p \to \mathbb{R}^{K_+}$ and $\boldsymbol{r}_{\mathrm{NN}} : \mathbb{R}^p \to \mathbb{R}^{K_-}$ are neural networks. Since the kernel $g(\boldsymbol{y},\boldsymbol{y}') = \langle \boldsymbol{y}_+, \boldsymbol{y}'_+\rangle - \langle \boldsymbol{y}_-, \boldsymbol{y}'_-\rangle$ with respect to $\boldsymbol{y} = (\boldsymbol{y}_+, \boldsymbol{y}_-) \in \mathbb{R}^{K_+ + K_-}$ represents the difference of two IPSs, we call (34) as inner product difference similarity (IPDS) model.

By replacing $\boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x})$ and $\boldsymbol{r}_{\mathrm{NN}}(\boldsymbol{x})$ with $(\boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x})^\top, u_{\mathrm{NN}}(\boldsymbol{x}), 1)^\top$ and $u_{\mathrm{NN}}(\boldsymbol{x}) - 1 \in \mathbb{R}$, respectively, IPDS reduces to SIPS defined in eq. (6), meaning that IPDS includes SIPS as a special case. Therefore, IPDS approximates any CPD similarities arbitrary well. Further, we prove that IPDS approximates more general similarities arbitrary well.

E.2 Approximation theorem

Theorem E.1 (Approximation theorem for IPDS) Symbols and assumptions are the same as those of Theorem 4.1 but g_* is a general kernel, which is only required to be dominated by some PD kernels g, i.e., $g - g_*$ is PD. For arbitrary $\varepsilon > 0$, by specifying sufficiently large $K_+, K_- \in \mathbb{N}, m_+ = m_+(K_+), m_- = m_-(K_-) \in \mathbb{N}$, there exist $\boldsymbol{A} \in \mathbb{R}^{K_+ \times m_+}, \boldsymbol{B} \in \mathbb{R}^{m_+ \times p}, \boldsymbol{c} \in \mathbb{R}^{m_+}, \boldsymbol{E} \in \mathbb{R}^{K_- \times m_-}, \boldsymbol{F} \in \mathbb{R}^{m_- \times p}, \boldsymbol{o} \in \mathbb{R}^{m_-}$ such that

$$\left| g_*\left(f_*(\boldsymbol{x}), f_*(\boldsymbol{x}')\right) - \left(\left\langle \boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x}), \boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x}')\right\rangle - \left\langle \boldsymbol{r}_{\mathrm{NN}}(\boldsymbol{x}), \boldsymbol{r}_{\mathrm{NN}}(\boldsymbol{x}')\right\rangle\right) \right| < \varepsilon$$

for all $(\boldsymbol{x},\boldsymbol{x}') \in [-M,M]^{2p}$, where $\boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{\sigma}(\boldsymbol{B}\boldsymbol{x}+\boldsymbol{c}) \in \mathbb{R}^{K_+}$ and $\boldsymbol{r}_{\mathrm{NN}}(x) = \boldsymbol{E}\boldsymbol{\sigma}(\boldsymbol{F}\boldsymbol{x}+\boldsymbol{o}) \in \mathbb{R}^{K_-}$ are 1-hidden layer neural networks with m_+ and m_- hidden units, respectively.

In theorem E.1, the kernel g_* is only required to be dominated by some PD kernels, thus g_* is not limited to CPD. We call such a kernel g_* satisfying the condition in Theorem E.1, i.e., there exists a PD kernel g such that $g - g_*$ is PD, as *general kernel*, and the general kernel g_* is called *indefinite* if neither of $g_*, -g_*$ is positive definite (Ong et al., 2004). General similarity and indefinite similarity are defined as well; IPDS approximates any general similarities arbitrary well.

Our proof for Theorem E.1 is based on Proposition 7 of Ong et al. (2004). This proposition indicates that the kernel g_* dominated by some PD kernels is decomposed as the difference of two PD kernels g_+, g_- by considering Krein space consisting of two Hilbert spaces. Therefore, we have $g_*(\boldsymbol{f}_*(\boldsymbol{x}), \boldsymbol{f}_*(\boldsymbol{x}')) = g_+(\boldsymbol{f}_*(\boldsymbol{x}), \boldsymbol{f}_*(\boldsymbol{x}')) - g_-(\boldsymbol{f}_*(\boldsymbol{x}), \boldsymbol{f}_*(\boldsymbol{x}'))$. Because of the PD-ness of g_+ and g_-, Theorem 3.2 guarantees the existence of NNs $\boldsymbol{f}_{\mathrm{NN}}, \boldsymbol{r}_{\mathrm{NN}}$ such that $\langle \boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x}), \boldsymbol{f}_{\mathrm{NN}}(\boldsymbol{x}')\rangle$ and $\langle \boldsymbol{r}_{\mathrm{NN}}(\boldsymbol{x}), \boldsymbol{r}_{\mathrm{NN}}(\boldsymbol{x}')\rangle$, respectively, approximate $g_+(\boldsymbol{f}_*(\boldsymbol{x}), \boldsymbol{f}_*(\boldsymbol{x}'))$ and $g_-(\boldsymbol{f}_*(\boldsymbol{x}), \boldsymbol{f}_*(\boldsymbol{x}'))$ arbitrary well. Thus proving the theorem. This idea for the proof is also interpreted as a generalized Mercer's theorem for Krein space (there is a similar attempt in Chen et al. (2008)) by applying Mercer's theorem to the two Hilbert spaces of Ong et al. (2004, Proposition 7).

E.3 Deep Gaussian embedding

To show another example of non-CPD kernels, Deep Gaussian embedding (Bojchevski and Günnemann, 2018) is reviewed below.

Example E.1 (Deep Gaussian embedding) Let $\mathcal{Y}$ be a set of distributions over a set $\boldsymbol{Z} \subset \mathbb{R}^q$. Kullback-Leibler divergence (Kullback and Leibler, 1951) between two distributions $\boldsymbol{y}, \boldsymbol{y}' \in \mathcal{Y}$ is defined by

$$d_{\mathrm{KL}}(\boldsymbol{y}, \boldsymbol{y}') := \int_{\boldsymbol{Z}} y(\boldsymbol{z}) \log \frac{y(\boldsymbol{z})}{y'(\boldsymbol{z})} \mathrm{d}\boldsymbol{z},$$

where $y(\boldsymbol{z})$ is the probability density function corresponding to the distribution $\boldsymbol{y} \in \mathcal{Y}$.

With the same setting in Section 2, Deep Gaussian embedding (Bojchevski and Günnemann, 2018), which incorporates neural networks into Gaussian embedding (Vilnis and McCallum, 2015), learns two neural networks $\boldsymbol{\mu} : \mathbb{R}^p \to \mathbb{R}^q, \boldsymbol{\Sigma} : \mathbb{R}^p \to \mathbb{R}_+^{q\times q}$ so that the function $\sigma(-d_{\mathrm{KL}}(\mathcal{N}_q(\boldsymbol{\mu}(\boldsymbol{x}_i), \boldsymbol{\Sigma}(\boldsymbol{x}_i)), \mathcal{N}_q(\boldsymbol{\mu}(\boldsymbol{x}_j), \boldsymbol{\Sigma}(\boldsymbol{x}_j))))$ approximates $E(w_{ij}|\boldsymbol{x}_i, \boldsymbol{x}_j)$. $\mathbb{R}_+^{q\times q}$ is a set of all $q \times q$ positive definite matrices and $\mathcal{N}_q(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ represents the q-variate normal distribution with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$.

Unlike typical graph embedding methods, deep Gaussian embedding maps data vectors to distributions as

$$\mathbb{R}^p \ni \boldsymbol{x} \mapsto \boldsymbol{y} := \mathcal{N}_q(\boldsymbol{\mu}(\boldsymbol{x}), \boldsymbol{\Sigma}(\boldsymbol{x})) \in \mathcal{Y},$$

where $\boldsymbol{y}$ is also interpreted as a vector of dimension $K = q + q(q+1)/2$ by considering the number of parameters in $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Our concern is to clarify if d_{KL} is CPD. However, in the first place, d_{KL} is not a kernel since it is not symmetric. In order to make it symmetric, Kullback-Leibler divergence may be replaced with Jeffrey's divergence (Kullback and Leibler, 1951)

$$d_{\mathrm{Jeff}}(\boldsymbol{y}, \boldsymbol{y}') := d_{\mathrm{KL}}(\boldsymbol{y}, \boldsymbol{y}') + d_{\mathrm{KL}}(\boldsymbol{y}', \boldsymbol{y}).$$

Although $-d_{\mathrm{Jeff}}$ is a kernel, it is not CPD as shown in Proposition E.1.

Proposition E.1 $-d_{\mathrm{Jeff}}$ is not CPD on $\tilde{\mathcal{P}}_q^2$, where $\tilde{\mathcal{P}}_q$ represents the set of all q-variate normal distributions.

A counterexample of CPD-ness is, $n = 3, q = 2, c_1 = -2/5, c_2 = -3/5, c_3 = 1, \boldsymbol{y}_i = \mathcal{N}_2(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \in \mathcal{Y}$ $(i = 1, 2, 3), \boldsymbol{\mu}_1 = (2, 1)^\top, \boldsymbol{\mu}_2 = (-1, 1)^\top, \boldsymbol{\mu}_3 = (1, 2)^\top, \boldsymbol{\Sigma}_1 = \mathrm{diag}(1/10, 1), \boldsymbol{\Sigma}_2 = \mathrm{diag}(1/2, 1), \boldsymbol{\Sigma}_3 = \mathrm{diag}(1, 1)$.

We are yet studying the nature of deep Gaussian embedding. However, as Proposition E.1 shows, negative Jeffrey's divergence used in the embedding is already proved to be non-CPD; SIPS cannot approximate it. IPDS model is required for approximating such non-CPD kernels. Thus we are currently trying to reveal to what extent IPDS applies, by classifying whether each of non-CPD kernels including negative Jeffrey's divergence satisfies the assumption on the kernel g_* in Theorem E.1.

References

Berg, C., Christensen, J., and Ressel, P. (1984). *Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions.* Graduate Texts in Mathematics. Springer New York.

Bojchevski, A. and Günnemann, S. (2018). Deep gaussian embedding of attributed graphs: Unsupervised inductive learning via ranking. In *Proceedings of the International Conference on Learning Representations (ICLR).*

Chen, D.-G., Wang, H.-Y., and Tsang, E. C. (2008). Generalized Mercer theorem and its application to feature space related to indefinite kernels. In *Proceedings of the International Conference on Machine Learning and Cybernetics (ICMLC)*, volume 2, pages 774–777. IEEE.

Cobos, F. and Kühn, T. (1990). Eigenvalues of integral operators with positive definite kernels satisfying integrated hölder conditions over metric compacta. *Journal of Approximation Theory*, 63(1):39–55.

Cybenko, G. (1989). Approximation by Superpositions of a Sigmoidal Function. *Mathematics of Control, Signals, and Systems (MCSS)*, 2(4):303–314.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Naber, G. L. (2012). *The geometry of Minkowski spacetime: An introduction to the mathematics of the special theory of relativity*, volume 92. Springer Science & Business Media.

Nickel, M. and Kiela, D. (2017). Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems (NIPS)*, pages 6341–6350.

Okuno, A., Hada, T., and Shimodaira, H. (2018). A probabilistic framework for multi-view feature learning with many-to-many associations via neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 3885–3894.

Ong, C. S., Mary, X., Canu, S., and Smola, A. J. (2004). Learning with non-positive kernels. In *Proceedings of the International Conference on Machine Learning (ICML)*, page 81. ACM.

Telgarsky, M. (2017). Neural networks and rational functions. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Vilnis, L. and McCallum, A. (2015). Word representations via gaussian embedding. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Yarotsky, D. (2018). Optimal approximation of continuous functions by very deep ReLU networks. *Proceedings of Machine Learning Research*, 75:639–649. The 31st Annual Conference on Learning Theory (COLT 2018).