Supplementary Material:
Robust Graph Embedding with Noisy Link Weights

A Lemmas and Proofs

A.1 Law of Large Numbers for Doubly-Indexed Partially-Dependent Random Variables

In this section, we first show and prove Theorem A.1, that is the law of large numbers, for doubly-indexed partially-dependent random variables. Then, we apply Theorem A.1 to the empirical probability β-score and the empirical moment β-score for proving Lemma A.2 and A.3 in which we show convergence

$$(6) \xrightarrow{p} E_{\mathcal{X}^2}(u_\beta^{(w_{12}|\boldsymbol{x}_1,\boldsymbol{x}_2)}(q,p_{\boldsymbol{\theta}})), \quad L_{\beta,n}(\boldsymbol{\theta}) \xrightarrow{p} u_\beta^{(\boldsymbol{x}_1,\boldsymbol{x}_2)}(g,\mu_{\boldsymbol{\theta}};\nu),$$

as $n \to \infty$, respectively.

Theorem A.1. Let $\boldsymbol{Z} := (Z_{ij})$ be an array of random variables $Z_{ij} \in \mathcal{Z}$, $(i,j) \in \mathcal{I}_n := \{(i,j) \mid 1 \le i < j \le n\}$, and $h : \mathcal{Z} \to \mathbb{R}$ be a bounded and continuous function. We assume that Z_{ij} is independent of Z_{kl} if $(k,l) \in \mathcal{R}_n(i,j) := \{(k,l) \in \mathcal{I}_n \mid k,l \in \{1,\dots,n\} \setminus \{i,j\}\}$, and $E_{\boldsymbol{Z}}(h(Z_{ij})^2) < \infty$, for all $(i,j) \in \mathcal{I}_n$. Then the average of $h(Z_{ij})$ over $\mathcal{I}_n$ converges to the expectation in probability as $n \to \infty$; that is

$$\frac{1}{|\mathcal{I}_n|}\sum_{(i,j)\in\mathcal{I}_n} h(Z_{ij}) = \frac{1}{|\mathcal{I}_n|}\sum_{(i,j)\in\mathcal{I}_n} E_{\boldsymbol{Z}}(h(Z_{ij})) + O_p(1/\sqrt{n}).$$

Proof of Theorem A.1. Regarding the variance of the average, we have

$$\begin{aligned}
V_{\boldsymbol{Z}}\left(\frac{1}{|\mathcal{I}_n|}\sum_{(i,j)\in\mathcal{I}_n} h(Z_{ij})\right) &= E_{\boldsymbol{Z}}\left(\left(\frac{1}{|\mathcal{I}_n|}\sum_{(i,j)\in\mathcal{I}_n} h(Z_{ij})\right)^2\right) - E_{\boldsymbol{Z}}\left(\frac{1}{|\mathcal{I}_n|}\sum_{(i,j)\in\mathcal{I}_n} h(Z_{ij})\right)^2 \\
&= \frac{1}{|\mathcal{I}_n|^2}\left(\sum_{(i,j)\in\mathcal{I}_n}\sum_{(k,l)\in\mathcal{I}_n} E_{\boldsymbol{Z}}\left(h(Z_{ij})h(Z_{kl})\right) - \left(\sum_{(i,j)\in\mathcal{I}_n} E_{\boldsymbol{Z}}\left(h(Z_{ij})\right)\right)^2\right) \\
&= \frac{1}{|\mathcal{I}_n|^2}\sum_{(i,j)\in\mathcal{I}_n}\sum_{(k,l)\in\mathcal{I}_n\setminus\mathcal{R}_n(i,j)} \left(E_{\boldsymbol{Z}}\left(h(Z_{ij})h(Z_{kl})\right) - E_{\boldsymbol{Z}}\left(h(Z_{ij})\right)E_{\boldsymbol{Z}}\left(h(Z_{kl})\right)\right),
\end{aligned}$$

where $E_{\boldsymbol{Z}}, V_{\boldsymbol{Z}}$ represent expectation and variance with respect to $\boldsymbol{Z}$. By considering $E_{\boldsymbol{Z}}(|h(Z_{ij})|) \le E_{\boldsymbol{Z}}(h(Z_{ij})^2)^{1/2} < \infty, E_{\boldsymbol{Z}}(|h(Z_{ij})h(Z_{kl})|) \le \sqrt{E_{\boldsymbol{Z}}(h(Z_{ij})^2)E_{\boldsymbol{Z}}(h(Z_{kl}))^2} < \infty$, $|\mathcal{I}_n| = O(n^2)$ and $|\mathcal{I}_n \setminus \mathcal{R}_n(i,j)| = O(n)$, the last formula is of order $O(n^{-4}\cdot n^2 \cdot n) = O(n^{-1})$. Therefore,

$$V_{\boldsymbol{Z}}\left(\frac{1}{|\mathcal{I}_n|}\sum_{(i,j)\in\mathcal{I}_n} h(Z_{ij})\right) = O(n^{-1}). \tag{21}$$

(21) and Chebyshev's inequality indicate the assertion. □

The same assertion appears in Supplement B.1 of Okuno et al. (2018). We note that the convergence rate is only $O_p(1/\sqrt{n})$ but not $O_p(1/\sqrt{|\mathcal{I}_n|}) = O_p(1/n)$, even though we leverage $O(|\mathcal{I}_n|) = O(n^2)$ observations $\{Z_{ij}\}_{(i,j)\in\mathcal{I}_n}$.

Lemma A.2. Let $\boldsymbol{\Theta}$ be a parameter set. Assuming that $w_{ij} \mid \boldsymbol{x}_i,\boldsymbol{x}_j \overset{\text{indep.}}{\sim} q$, $\boldsymbol{x}_i \overset{\text{i.i.d.}}{\sim} Q$, $\mathrm{supp}Q \subset \mathcal{X}$ where $\mathcal{X} \subset \mathbb{R}^p$ is a compact set, $\sum_{w\in\mathbb{N}_0} q(w \mid \boldsymbol{x}_1,\boldsymbol{x}_2)p_{\boldsymbol{\theta}}(w \mid \boldsymbol{x}_1,\boldsymbol{x}_2)^\delta < \infty, \sum_{w\in\mathbb{N}_0} p_{\boldsymbol{\theta}}(w \mid \boldsymbol{x}_1,\boldsymbol{x}_2)^{1+\delta} < \infty$ for all $\delta > 0, \boldsymbol{x}_1,\boldsymbol{x}_2 \in \mathcal{X}$. Then, it holds for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ that

$$(6) = E_{\mathcal{X}^2}(d_\beta^{(w_{12}|\boldsymbol{x}_1,\boldsymbol{x}_2)}(q,p_{\boldsymbol{\theta}})) + O_p(1/\sqrt{n}),$$

indicating $(6) \xrightarrow{p} E_{\mathcal{X}^2}(d_\beta^{(w_{12}|\boldsymbol{x}_1,\boldsymbol{x}_2)}(q,p_{\boldsymbol{\theta}}))$ $(n \to \infty)$.

Proof of Lemma A.2. Applying Theorem A.1 to

$$Z_{ij} := (w_{ij}, \boldsymbol{x}_i, \boldsymbol{x}_j),\ h(Z_{ij}) := -\frac{p_{\boldsymbol{\theta}}(w_{ij} \mid \boldsymbol{x}_i, \boldsymbol{x}_j)^\beta - 1}{\beta} + \sum_{w \in \mathbb{N}_0} \frac{p_{\boldsymbol{\theta}}(w \mid \boldsymbol{x}_i, \boldsymbol{x}_j)^{1+\beta}}{1+\beta},$$

immediately proves the assertion, as $E_{\boldsymbol{Z}}(h(Z_{ij})^2) < \infty$ follows from the assumptions; the convergence limit is,

$$\begin{aligned}
&\frac{1}{|\mathcal{I}_n|} \sum_{(i,j) \in \mathcal{I}_n} E_{\boldsymbol{Z}}(h(Z_{ij})) \\
&= \frac{1}{|\mathcal{I}_n|} \sum_{(i,j) \in \mathcal{I}_n} E_{\mathcal{X}^2} \left(E \left(-\frac{p_{\boldsymbol{\theta}}(w_{ij} \mid \boldsymbol{x}_i, \boldsymbol{x}_j)^\beta - 1}{\beta} + \sum_{w \in \mathbb{N}_0} \frac{p_{\boldsymbol{\theta}}(w \mid \boldsymbol{x}_i, \boldsymbol{x}_j)^{1+\beta}}{1+\beta} \middle| \boldsymbol{x}_i, \boldsymbol{x}_j \right) \right) \\
&= \frac{1}{|\mathcal{I}_n|} \sum_{(i,j) \in \mathcal{I}_n} E_{\mathcal{X}^2} \left(\sum_{w' \in \mathbb{N}_0} q(w' \mid \boldsymbol{x}_i, \boldsymbol{x}_j) \left\{ -\frac{p_{\boldsymbol{\theta}}(w' \mid \boldsymbol{x}_i, \boldsymbol{x}_j)^\beta - 1}{\beta} + \sum_{w \in \mathbb{N}_0} \frac{p_{\boldsymbol{\theta}}(w \mid \boldsymbol{x}_i, \boldsymbol{x}_j)^{1+\beta}}{1+\beta} \right\} \right) \\
&= \frac{1}{|\mathcal{I}_n|} \sum_{(i,j) \in \mathcal{I}_n} E_{\mathcal{X}^2} \left(-\sum_{w \in \mathbb{N}_0} q(w \mid \boldsymbol{x}_i, \boldsymbol{x}_j) \frac{p_{\boldsymbol{\theta}}(w \mid \boldsymbol{x}_i, \boldsymbol{x}_j)^\beta - 1}{\beta} + \sum_{w \in \mathbb{N}_0} \frac{p_{\boldsymbol{\theta}}(w \mid \boldsymbol{x}_i, \boldsymbol{x}_j)^{1+\beta}}{1+\beta} \right) \\
&= \frac{1}{|\mathcal{I}_n|} \sum_{(i,j) \in \mathcal{I}_n} E_{\mathcal{X}^2}(d_\beta^{(w_{ij} \mid \boldsymbol{x}_i, \boldsymbol{x}_j)}(q, p_{\boldsymbol{\theta}})) \\
&= \frac{1}{|\mathcal{I}_n|} \sum_{(i,j) \in \mathcal{I}_n} E_{\mathcal{X}^2}(d_\beta^{(w_{12} \mid \boldsymbol{x}_1, \boldsymbol{x}_2)}(q, p_{\boldsymbol{\theta}})) \\
&= E_{\mathcal{X}^2}(d_\beta^{(w_{12} \mid \boldsymbol{x}_1, \boldsymbol{x}_2)}(q, p_{\boldsymbol{\theta}})).
\end{aligned}$$

Thus proving the assertion. □

Lemma A.3. Let $\boldsymbol{\Theta}$ be a parameter set. Assuming (9)–(11), it holds for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ that

$$L_{\beta,n}(\boldsymbol{\theta}) = u_\beta^{(\boldsymbol{x}_1, \boldsymbol{x}_2)}(g, \mu_{\boldsymbol{\theta}}; \nu) + O_p(1/\sqrt{n}),$$

indicating $L_{\beta,n}(\boldsymbol{\theta}) \xrightarrow{p} u_\beta^{(\boldsymbol{x}_1, \boldsymbol{x}_2)}(g, \mu_{\boldsymbol{\theta}}; \nu)$ $(n \to \infty)$.

Proof of Lemma A.3. Applying Theorem A.1 to

$$Z_{ij} := (w_{ij}, \boldsymbol{x}_i, \boldsymbol{x}_j),\ h(Z_{ij}) := -w_{ij} \frac{\mu_{\boldsymbol{\theta}}(\boldsymbol{x}_i, \boldsymbol{x}_j)^\beta - 1}{\beta} + \frac{\mu_{\boldsymbol{\theta}}(\boldsymbol{x}_i, \boldsymbol{x}_j)^{1+\beta}}{1+\beta},$$

immediately proves the assertion, as $E_{\boldsymbol{Z}}(h(Z_{ij})^2) < \infty$ follows from the assumptions; the convergence limit is,

$$\begin{aligned}
\frac{1}{|\mathcal{I}_n|} \sum_{(i,j) \in \mathcal{I}_n} E_{\boldsymbol{Z}}(h(Z_{ij})) &= \frac{1}{|\mathcal{I}_n|} \sum_{(i,j) \in \mathcal{I}_n} E_{\mathcal{X}^2} \left(E \left(-w_{ij} \frac{\mu_{\boldsymbol{\theta}}(\boldsymbol{x}_i, \boldsymbol{x}_j)^\beta - 1}{\beta} + \frac{\mu_{\boldsymbol{\theta}}(\boldsymbol{x}_i, \boldsymbol{x}_j)^{1+\beta}}{1+\beta} \middle| \boldsymbol{x}_i, \boldsymbol{x}_j \right) \right) \\
&= \frac{1}{|\mathcal{I}_n|} \sum_{(i,j) \in \mathcal{I}_n} E_{\mathcal{X}^2} \left(-g(\boldsymbol{x}_i, \boldsymbol{x}_j) \frac{\mu_{\boldsymbol{\theta}}(\boldsymbol{x}_i, \boldsymbol{x}_j)^\beta - 1}{\beta} + \frac{\mu_{\boldsymbol{\theta}}(\boldsymbol{x}_i, \boldsymbol{x}_j)^{1+\beta}}{1+\beta} \right) \\
&= \frac{1}{|\mathcal{I}_n|} \sum_{(i,j) \in \mathcal{I}_n} u_\beta^{(\boldsymbol{x}_i, \boldsymbol{x}_j)}(g, \mu_{\boldsymbol{\theta}}; \nu) \\
&= \frac{1}{|\mathcal{I}_n|} \sum_{(i,j) \in \mathcal{I}_n} u_\beta^{(\boldsymbol{x}_1, \boldsymbol{x}_2)}(g, \mu_{\boldsymbol{\theta}}; \nu) \\
&= u_\beta^{(\boldsymbol{x}_1, \boldsymbol{x}_2)}(g, \mu_{\boldsymbol{\theta}}; \nu).
\end{aligned}$$

Thus proving the assertion. □

A.2 Evaluation of $M(\boldsymbol{\theta})$ in Theorem 3.1

Lemma A.4. Suppose that $\varepsilon \geq \varepsilon_*$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}_\varepsilon := \{\boldsymbol{\theta} \in \boldsymbol{\Theta} \mid E_{\mathcal{X}^2}(\eta_*(\boldsymbol{x}_1,\boldsymbol{x}_2)\mu_{\boldsymbol{\theta}}(\boldsymbol{x}_1,\boldsymbol{x}_2)^{\beta_0}) < \varepsilon\}$, and $\beta \in (0,\beta_0]$, it holds for

$$M(\boldsymbol{\theta}) := \beta^{-1} E_{\mathcal{X}^2}\left(\eta_*(\boldsymbol{x}_1,\boldsymbol{x}_2)\mu_{\boldsymbol{\theta}}(\boldsymbol{x}_1,\boldsymbol{x}_2)^{\beta}\right)\varepsilon^{-\beta/\beta_0}, \quad \alpha := E_{\mathcal{X}^2}(\eta_*(\boldsymbol{x}_1,\boldsymbol{x}_2)),$$

that

$$M(\boldsymbol{\theta}) \leq \alpha^{1-\beta/\beta_0}\beta^{-1} \quad (\forall \boldsymbol{\theta} \in \boldsymbol{\Theta}_\varepsilon).$$

Proof of Lemma A.4. Proof is based on Lyapunov's inequality, that is, $E(Z^\beta) \leq E(Z^{\beta_0})^{\beta/\beta_0}$ for any non-negative real-valued random variable Z and $0 < \beta \leq \beta_0 < \infty$. For applying this inequality, we first fix $\boldsymbol{\theta} \in \boldsymbol{\Theta}_\varepsilon$, and expand $M(\boldsymbol{\theta})$ with the probability density function (pdf) ν of the random variable $(\boldsymbol{x}_1,\boldsymbol{x}_2)$ as

$$\begin{aligned} M(\boldsymbol{\theta}) &= \beta^{-1} E_{\mathcal{X}^2}\left(\eta_*(\boldsymbol{x}_1,\boldsymbol{x}_2)\mu_{\boldsymbol{\theta}}(\boldsymbol{x}_1,\boldsymbol{x}_2)^{\beta}\right)\varepsilon^{-\beta/\beta_0} \\ &= \beta^{-1}\varepsilon^{-\beta/\beta_0} \iint_{\mathcal{X}^2} \nu(\boldsymbol{x}_1,\boldsymbol{x}_2)\eta_*(\boldsymbol{x}_1,\boldsymbol{x}_2)\mu_{\boldsymbol{\theta}}(\boldsymbol{x}_1,\boldsymbol{x}_2)^\beta \mathrm{d}\boldsymbol{x}_1\mathrm{d}\boldsymbol{x}_2 \\ &= \alpha\beta^{-1}\varepsilon^{-\beta/\beta_0}\left(\iint_{\mathcal{X}^2} \underbrace{\frac{\nu(\boldsymbol{x}_1,\boldsymbol{x}_2)\eta_*(\boldsymbol{x}_1,\boldsymbol{x}_2)}{\alpha}}_{=:\tilde{\nu}(\boldsymbol{x}_1,\boldsymbol{x}_2)} \mu_{\boldsymbol{\theta}}(\boldsymbol{x}_1,\boldsymbol{x}_2)^\beta \mathrm{d}\boldsymbol{x}_1\mathrm{d}\boldsymbol{x}_2\right). \end{aligned} \tag{22}$$

In eq. (22), $\tilde{\nu}(\boldsymbol{x}_1,\boldsymbol{x}_2) := \nu(\boldsymbol{x}_1,\boldsymbol{x}_2)\eta_*(\boldsymbol{x}_1,\boldsymbol{x}_2)/\alpha$ can be regarded as a pdf, since $\tilde{\nu}(\boldsymbol{x}_1,\boldsymbol{x}_2) \geq 0$ for all $(\boldsymbol{x}_1,\boldsymbol{x}_2)$ and

$$\begin{aligned} \iint_{\mathcal{X}^2} \tilde{\nu}_*(\boldsymbol{x}_1,\boldsymbol{x}_2)\mathrm{d}\boldsymbol{x}_1\mathrm{d}\boldsymbol{x}_2 &= \iint_{\mathcal{X}^2} \frac{\nu(\boldsymbol{x}_1,\boldsymbol{x}_2)\eta_*(\boldsymbol{x}_1,\boldsymbol{x}_2)}{\alpha}\mathrm{d}\boldsymbol{x}_1\mathrm{d}\boldsymbol{x}_2 \\ &= \alpha^{-1}\iint_{\mathcal{X}^2} \nu(\boldsymbol{x}_1,\boldsymbol{x}_2)\eta_*(\boldsymbol{x}_1,\boldsymbol{x}_2)\mathrm{d}\boldsymbol{x}_1\mathrm{d}\boldsymbol{x}_2 \\ &= \alpha^{-1}E_{\mathcal{X}^2}(\eta_*(\boldsymbol{x}_1,\boldsymbol{x}_2)) = \alpha^{-1}\alpha = 1. \end{aligned}$$

As $\tilde{\nu}$ can be regarded as a pdf and $\mu_{\boldsymbol{\theta}}$ is non-negative, Lyapunov's inequality indicates that

$$\begin{aligned} M(\boldsymbol{\theta}) = (22) &\overset{\text{(Lyapunov)}}{\leq} \alpha\beta^{-1}\varepsilon^{-\beta/\beta_0}\left(\iint_{\mathcal{X}^2} \tilde{\nu}(\boldsymbol{x}_1,\boldsymbol{x}_2)\mu_{\boldsymbol{\theta}}(\boldsymbol{x}_1,\boldsymbol{x}_2)^{\beta_0}\mathrm{d}\boldsymbol{x}_1\mathrm{d}\boldsymbol{x}_2\right)^{\beta/\beta_0} \\ &= \alpha\beta^{-1}\varepsilon^{-\beta/\beta_0}\left(\iint_{\mathcal{X}^2} \frac{\nu(\boldsymbol{x}_1,\boldsymbol{x}_2)\eta_*(\boldsymbol{x}_1,\boldsymbol{x}_2)}{\alpha}\mu_{\boldsymbol{\theta}}(\boldsymbol{x}_1,\boldsymbol{x}_2)^{\beta_0}\mathrm{d}\boldsymbol{x}_1\mathrm{d}\boldsymbol{x}_2\right)^{\beta/\beta_0} \\ &= \alpha^{1-\beta/\beta_0}\beta^{-1}\varepsilon^{-\beta/\beta_0}\left(\iint_{\mathcal{X}^2} \nu(\boldsymbol{x}_1,\boldsymbol{x}_2)\eta_*(\boldsymbol{x}_1,\boldsymbol{x}_2)\mu_{\boldsymbol{\theta}}(\boldsymbol{x}_1,\boldsymbol{x}_2)^{\beta_0}\mathrm{d}\boldsymbol{x}_1\mathrm{d}\boldsymbol{x}_2\right)^{\beta/\beta_0} \\ &= \alpha^{1-\beta/\beta_0}\beta^{-1}\varepsilon^{-\beta/\beta_0}E_{\mathcal{X}^2}\left(\eta_*(\boldsymbol{x}_1,\boldsymbol{x}_2)\mu_{\boldsymbol{\theta}}(\boldsymbol{x}_1,\boldsymbol{x}_2)^{\beta_0}\right)^{\beta/\beta_0} \\ &\leq \alpha^{1-\beta/\beta_0}\beta^{-1}\varepsilon^{-\beta/\beta_0}\varepsilon^{\beta/\beta_0} \qquad (\because \boldsymbol{\theta} \in \boldsymbol{\Theta}_\varepsilon) \\ &= \alpha^{1-\beta/\beta_0}\beta^{-1}. \end{aligned}$$

The assertion is proved. □

A.3 Proof of Theorem 3.2

We first verify that (19) is equivalent to $\partial h(\boldsymbol{\theta})/\partial\boldsymbol{\theta} = \boldsymbol{0}$. From the definition of $h^{(t)}(\boldsymbol{\theta})$ and the assumption (i) $\mu_{\boldsymbol{\theta}}(\boldsymbol{x}_1,\boldsymbol{x}_2) \in C^1(\boldsymbol{\Theta})$ for all $(\boldsymbol{x}_1,\boldsymbol{x}_2) \in \mathcal{X}^2$, we have

$$\begin{aligned} \frac{\partial h(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} &= \frac{\partial E^{(1)}(h^{(1)}(\boldsymbol{\theta}))}{\partial\boldsymbol{\theta}} = E^{(1)}\left(\frac{\partial h^{(1)}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right) \\ &= E^{(1)}\left(\frac{\partial}{\partial\boldsymbol{\theta}}\left\{-\sum_{(i,j)\in\mathcal{W}_n^{(1)}} w_{ij}\frac{\mu_{\boldsymbol{\theta}}(\boldsymbol{x}_i,\boldsymbol{x}_j)^\beta - 1}{\beta} + \lambda\sum_{(i,j)\in\mathcal{I}_n^{(1)}}\frac{\mu_{\boldsymbol{\theta}}(\boldsymbol{x}_i,\boldsymbol{x}_j)^{1+\beta}}{1+\beta}\right\}\right) \end{aligned}$$

$$
\begin{aligned}
&= E^{(1)}\left(\left\{-\sum_{(i,j)\in\mathcal{W}_n^{(1)}} w_{ij}\mu_{\boldsymbol{\theta}}(\boldsymbol{x}_i,\boldsymbol{x}_j)^{\beta}\frac{\partial\log\mu_{\boldsymbol{\theta}}(\boldsymbol{x}_i,\boldsymbol{x}_j)}{\partial\boldsymbol{\theta}} + \lambda\sum_{(i,j)\in\mathcal{I}_n^{(1)}}\mu_{\boldsymbol{\theta}}(\boldsymbol{x}_i,\boldsymbol{x}_j)^{1+\beta}\frac{\partial\log\mu_{\boldsymbol{\theta}}(\boldsymbol{x}_i,\boldsymbol{x}_j)}{\partial\boldsymbol{\theta}}\right\}\right) \\
&= -E^{(1)}\left(\sum_{(i,j)\in\mathcal{W}_n^{(1)}} w_{ij}\mu_{\boldsymbol{\theta}}(\boldsymbol{x}_i,\boldsymbol{x}_j)^{\beta}\frac{\partial\log\mu_{\boldsymbol{\theta}}(\boldsymbol{x}_i,\boldsymbol{x}_j)}{\partial\boldsymbol{\theta}}\right) + \lambda E^{(1)}\left(\sum_{(i,j)\in\mathcal{I}_n^{(1)}}\mu_{\boldsymbol{\theta}}(\boldsymbol{x}_i,\boldsymbol{x}_j)^{1+\beta}\frac{\partial\log\mu_{\boldsymbol{\theta}}(\boldsymbol{x}_i,\boldsymbol{x}_j)}{\partial\boldsymbol{\theta}}\right) \\
&= -\frac{m_1}{|\mathcal{W}_n|}\sum_{(i,j)\in\mathcal{W}_n} w_{ij}\mu_{\boldsymbol{\theta}}(\boldsymbol{x}_i,\boldsymbol{x}_j)^{\beta}\frac{\partial\log\mu_{\boldsymbol{\theta}}(\boldsymbol{x}_i,\boldsymbol{x}_j)}{\partial\boldsymbol{\theta}} + \lambda\frac{m_2}{|\mathcal{I}_n|}\sum_{(i,j)\in\mathcal{I}_n}\mu_{\boldsymbol{\theta}}(\boldsymbol{x}_i,\boldsymbol{x}_j)^{1+\beta}\frac{\partial\log\mu_{\boldsymbol{\theta}}(\boldsymbol{x}_i,\boldsymbol{x}_j)}{\partial\boldsymbol{\theta}} \\
&= \frac{1}{|\mathcal{I}_n|}\sum_{(i,j)\in\mathcal{I}_n}\left\{\left(-vm_1w_{ij} + \lambda m_2\mu_{\boldsymbol{\theta}}(\boldsymbol{x}_i,\boldsymbol{x}_j)\right)\mu_{\boldsymbol{\theta}}(\boldsymbol{x}_i,\boldsymbol{x}_j)^{\beta}\frac{\partial\log\mu_{\boldsymbol{\theta}}(\boldsymbol{x}_i,\boldsymbol{x}_j)}{\partial\boldsymbol{\theta}}\right\}.
\end{aligned}
$$

We next verify the convergence $E^*(\|\boldsymbol{\theta}^{(t)}-\boldsymbol{\theta}_*\|_2^2)\to 0$. From the assumption (ii), $\boldsymbol{\theta}_*$ is the unique minimizer of $h(\boldsymbol{\theta})$ over $\boldsymbol{\Theta}$. Regarding the estimator $\boldsymbol{\theta}^{(t)}$ defined as (18) with the assumption (iii), Moulines and Bach (2011) Theorem 2 asserts that $E^*(\|\boldsymbol{\theta}^{(t)}-\boldsymbol{\theta}_*\|_2^2)\to 0$ if the following conditions (C-1)–(C-3) hold: (C-1) $E^{(t)}\left(\frac{\partial h^{(t)}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right) = \frac{\partial h(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}$ for all $\boldsymbol{\theta}\in\boldsymbol{\Theta}$, (C-2) $h(\boldsymbol{\theta})$ is strongly convex on $\boldsymbol{\Theta}$, i.e., $\exists\lambda>0$ such that $h(\boldsymbol{\theta}_1)-h(\boldsymbol{\theta}_2)\geq\langle\frac{\partial h(\boldsymbol{\theta}_2)}{\partial\boldsymbol{\theta}},\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\rangle+\lambda\|\boldsymbol{\theta}_1-\boldsymbol{\theta}_2\|_2^2$ for all $\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\in\boldsymbol{\Theta}$, and (C-3) $\|\frac{\partial h^{(t)}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\|_2$ is bounded on $\boldsymbol{\Theta}$ for any $(\mathcal{W}_n^{(t)},\mathcal{I}_n^{(t)})$. These conditions (C-1)–(C-3) correspond to the conditions (H1), (H3), and (H5), that are required in Moulines and Bach (2011) Theorem 2, respectively.

In case of Theorem 3.2, (C-1) holds as we have already seen for showing (19); note that $h^{(t)}(\boldsymbol{\theta})\in C^1(\boldsymbol{\Theta})$ from the assumption (i). (C-2) is assumed as (ii), and (C-3) holds because $h^{(t)}(\boldsymbol{\theta})$ is C^1 on the compact set $\boldsymbol{\Theta}$ and $(\mathcal{W}_n^{(t)},\mathcal{I}_n^{(t)})$ is a random variable taking value in a finite set. Thus we have proved the convergence.

□

References

Moulines, E. and Bach, F. R. (2011). Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems (NIPS)*, pages 451–459.

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)*.