Supplements

A Proofs

A.1 Proof of Theorem 1

We first introduce the following technical lemma.

Lemma 3. Let $g(\boldsymbol{\theta})$, $f(\boldsymbol{\theta})$, and $h(\boldsymbol{\theta})$ be defined as in Section 2.1; hence $f(\boldsymbol{\theta})$ is convex and differentiable, and $\boldsymbol{\nabla} f(\boldsymbol{\theta})$ is Lipschitz continuous with Lipschitz constant L. Let $\alpha \leq 1/L$. Let $\boldsymbol{G}_\alpha(\boldsymbol{\theta})$ and $\boldsymbol{\Delta} f(\boldsymbol{\theta})$ be defined as in Section (2.2). Then for all $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, the following inequality holds:

$$g(\boldsymbol{\theta}_1^\dagger) \leq g(\boldsymbol{\theta}_2) + \boldsymbol{G}_\alpha^\top(\boldsymbol{\theta}_1)(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) + (\boldsymbol{\nabla} f(\boldsymbol{\theta}_1) - \boldsymbol{\Delta} f(\boldsymbol{\theta}_1))^\top (\boldsymbol{\theta}_1^\dagger - \boldsymbol{\theta}_2) - \frac{\alpha}{2} \left\| \boldsymbol{G}_\alpha(\boldsymbol{\theta}_1) \right\|_2^2, \tag{16}$$

where $\boldsymbol{\theta}_1^\dagger = \boldsymbol{\theta}_1 - \alpha \boldsymbol{G}_\alpha(\boldsymbol{\theta}_1)$.

Proof. The proof is based on the convergence analysis of the standard proximal gradient method [Vandenberghe, 2016]. $f(\boldsymbol{\theta})$ is a convex differentiable function whose gradient is Lipschitz continuous with Lipschitz constant L. By the quadratic bound of the Lipschitz property:

$$f(\boldsymbol{\theta}_1^\dagger) \leq f(\boldsymbol{\theta}_1) - \alpha \boldsymbol{\nabla}^\top f(\boldsymbol{\theta}_1) \boldsymbol{G}_\alpha(\boldsymbol{\theta}_1) + \frac{\alpha^2 L}{2} \left\| \boldsymbol{G}_\alpha(\boldsymbol{\theta}_1) \right\|_2^2.$$

With $\alpha \leq 1/L$, and adding $h(\boldsymbol{\theta}_1^\dagger)$ on both sides of the quadratic bound, we have an upper bound for $g(\boldsymbol{\theta}_1^\dagger)$:

$$g(\boldsymbol{\theta}_1^\dagger) \leq f(\boldsymbol{\theta}_1) - \alpha \boldsymbol{\nabla}^\top f(\boldsymbol{\theta}_1) \boldsymbol{G}_\alpha(\boldsymbol{\theta}_1) + \frac{\alpha}{2} \left\| \boldsymbol{G}_\alpha(\boldsymbol{\theta}_1) \right\|_2^2 + h(\boldsymbol{\theta}_1^\dagger).$$

By convexity of $f(\boldsymbol{\theta})$ and $h(\boldsymbol{\theta})$, we have:

$$f(\boldsymbol{\theta}_1) \leq f(\boldsymbol{\theta}_2) + \boldsymbol{\nabla}^\top f(\boldsymbol{\theta}_1)(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2),$$
$$h(\boldsymbol{\theta}_1^\dagger) \leq h(\boldsymbol{\theta}_2) + (\boldsymbol{G}_\alpha(\boldsymbol{\theta}_1) - \Delta f(\boldsymbol{\theta}_1))^\top (\boldsymbol{\theta}_1^+ - \boldsymbol{\theta}_2),$$

which can be used to further upper bound $g(\boldsymbol{\theta}_1^\dagger)$, and results in (16). Note that we have used the fact that $G_\alpha(\boldsymbol{\theta}_1) - \Delta f(\boldsymbol{\theta}_1)$ is a subgradient of $h(\boldsymbol{\theta}_1^\dagger)$ in the last inequality. □

With Lemma 3, we are now able to prove Theorem 1. In Lemma 3, let $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 = \boldsymbol{\theta}^{(k)}$. Then by (8), $\boldsymbol{\theta}_1^\dagger = \boldsymbol{\theta}^{(k+1)}$. The inequality in (16) can then be simplified as:

$$g(\boldsymbol{\theta}^{(k+1)}) - g(\boldsymbol{\theta}^{(k)}) \leq \alpha \boldsymbol{\delta}(\boldsymbol{\theta}^{(k)})^\top \boldsymbol{G}_\alpha(\boldsymbol{\theta}^{(k)}) - \frac{\alpha}{2} \|\boldsymbol{G}_\alpha(\boldsymbol{\theta}^{(k)})\|_2^2.$$

By the Cauchy-Schwarz inequality and the sufficient condition that $\|\boldsymbol{\delta}(\boldsymbol{\theta}^{(k)})\|_2 < \frac{1}{2}\|\boldsymbol{G}_\alpha(\boldsymbol{\theta}^{(k)})\|_2$, we can further simplify the inequality and conclude $g(\boldsymbol{\theta}^{(k+1)}) < g(\boldsymbol{\theta}^{(k)})$.

A.2 Proof of Theorem 2

To prove Theorem 2, we first review Proposition 1 in Schmidt et al. [2011]:

Theorem 7 (Convergence on Average, Schmidt et al. [2011])**.** Let $\mathcal{K} = (\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \cdots, \boldsymbol{\theta}^{(\kappa)})$ be the iterates generated by Algorithm 3, then

$$g\left(\frac{1}{\kappa}\sum_{k=1}^{\kappa} \boldsymbol{\theta}^{(k)}\right) - g(\hat{\boldsymbol{\theta}}) \leq \frac{L}{2\kappa}\left(\|\boldsymbol{\theta}^{(0)} - \hat{\boldsymbol{\theta}}\|_2 + \frac{2}{L}\sum_{k=1}^{\kappa}\|\boldsymbol{\delta}(\boldsymbol{\theta}^{(k)})\|_2\right)^2.$$

Furthermore, according to the assumption that $g(\boldsymbol{\theta}^{(k+1)}) \leq g(\boldsymbol{\theta}^{(k)})$ with $k \in \{1, 2, \cdots, \kappa\}$, we have: $g\left(\frac{1}{\kappa}\sum_{k=1}^{\kappa} \boldsymbol{\theta}^{(k)}\right) \geq g(\boldsymbol{\theta}^{(\kappa)})$. Therefore,

$$g(\boldsymbol{\theta}^{(\kappa)}) - g(\hat{\boldsymbol{\theta}}) \leq \frac{L}{2\kappa}\left(\|\boldsymbol{\theta}^{(0)} - \hat{\boldsymbol{\theta}}\|_2 + \frac{2}{L}\sum_{k=1}^{\kappa}\|\boldsymbol{\delta}(\boldsymbol{\theta}^{(k)})\|_2\right)^2.$$

A.3 Proof of Theorem 3

A.3.1 Proof of Lemma 1

The rationale behind our proof follow that of Bengio and Delalleau [2009] and Fischer and Igel [2011].

Let $\tilde{\mathbf{x}}_0 \in \{0,1\}^p$ be an initialization of the Gibbs sampling algorithm. Let $\boldsymbol{\theta}$ be the parameterization from which the Gibbs sampling algorithm generates new samples. A Gibbs-τ algorithm hence uses the τ^{th} sample , $\tilde{\mathbf{x}}_\tau$, generated from the chain to approximate the gradient. Since there is only one Markov chain in total, we have $\mathbb{S} = \{\tilde{\mathbf{x}}_\tau\}$. The gradient approximation of Gibbs-τ is hence given by:

$$\boldsymbol{\Delta} f(\boldsymbol{\theta}) = \boldsymbol{\psi}(\tilde{\mathbf{x}}_\tau) - \mathbb{E}_{\mathbb{X}}\boldsymbol{\psi}(\mathbf{x}). \tag{17}$$

The actual gradient, $\boldsymbol{\nabla} f(\boldsymbol{\theta})$, is given in (3). Therefore, the difference between the approximation and the actual gradient is

$$\boldsymbol{\delta}(\boldsymbol{\theta}) = \boldsymbol{\Delta} f(\boldsymbol{\theta}) - \boldsymbol{\nabla} f(\boldsymbol{\theta}) = \boldsymbol{\psi}(\tilde{\mathbf{x}}_\tau) - \mathbb{E}_{\boldsymbol{\theta}}\boldsymbol{\psi}(\mathbf{x}) = \boldsymbol{\nabla} \log \mathrm{P}_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}_\tau).$$

We rewrite

$$\mathrm{P}_\tau(\mathbf{x} \mid \tilde{\mathbf{x}}_0) = \mathrm{P}(\tilde{\mathbf{X}}_\tau = \mathbf{x} \mid \tilde{\mathbf{x}}_0) = \mathrm{P}_{\boldsymbol{\theta}}(\mathbf{x}) + \epsilon_\tau(\mathbf{x}),$$

where $\epsilon_\tau(\mathbf{x})$ is the difference between $\mathrm{P}_\tau(\mathbf{x} \mid \tilde{\mathbf{x}}_0)$ and $\mathrm{P}_{\boldsymbol{\theta}}(\mathbf{x})$. Consider the expectation of the j^{th} component of $\boldsymbol{\delta}(\boldsymbol{\theta})$, $\delta_j(\boldsymbol{\theta})$, where $j \in \{1, 2, \cdots, m\}$, after running Gibbs-$\tau$ that is initialized by $\tilde{\mathbf{x}}_0$:

$$\begin{aligned}
\mathbb{E}_{\tilde{\mathbf{x}}_\tau}[\delta_j(\boldsymbol{\theta}) \mid \tilde{\mathbf{x}}_0] &= \sum_{\mathbf{x}\in\{0,1\}^p} \mathrm{P}_\tau(\mathbf{x} \mid \tilde{\mathbf{x}}_0)\delta_j(\boldsymbol{\theta}) = \sum_{\mathbf{x}\in\{0,1\}^p} (\mathrm{P}_{\boldsymbol{\theta}}(\mathbf{x}) + \epsilon_\tau(\mathbf{x}))\delta_j(\boldsymbol{\theta}) \\
&= \sum_{\mathbf{x}\in\{0,1\}^p} \epsilon_\tau(\mathbf{x})\delta_i(\boldsymbol{\theta}) = \sum_{\mathbf{x}\in\{0,1\}^p} (\mathrm{P}_\tau(\mathbf{x} \mid \mathbf{x}_0) - \mathrm{P}_{\boldsymbol{\theta}}(\mathbf{x}))\delta_j(\boldsymbol{\theta}) \\
&= \sum_{\mathbf{x}\in\{0,1\}^p} (\mathrm{P}_\tau(\mathbf{x} \mid \mathbf{x}_0) - \mathrm{P}_{\boldsymbol{\theta}}(\mathbf{x}))\nabla_j \log \mathrm{P}_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}_\tau),
\end{aligned} \tag{18}$$

where we have used the fact that $\sum_{\mathbf{x}\in\{0,1\}^p} \mathrm{P}_{\boldsymbol{\theta}}(\mathbf{x})\nabla_j \log \mathrm{P}_{\boldsymbol{\theta}}(\mathbf{x}) = 0$, and $\nabla_j \log \mathrm{P}_{\boldsymbol{\theta}}(\mathbf{x})$ represents the j^{th} component of $\boldsymbol{\nabla} \log \mathrm{P}_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}_\tau)$, with $j \in \{1, 2, \cdots, m\}$.

Therefore, from (18),

$$\begin{aligned}
|\mathbb{E}_{\tilde{\mathbf{x}}_\tau}[\delta_j(\boldsymbol{\theta}) \mid \tilde{\mathbf{x}}_0]| &\le \sum_{\mathbf{x}\in\{0,1\}^p} |\mathrm{P}_\tau(\mathbf{x} \mid \mathbf{x}_0) - \mathrm{P}_{\boldsymbol{\theta}}(\mathbf{x})| \cdot |\nabla_j \log \mathrm{P}_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}_\tau)| \\
&\le \sum_{\mathbf{x}\in\{0,1\}^p} |\mathrm{P}_\tau(\mathbf{x} \mid \mathbf{x}_0) - \mathrm{P}_{\boldsymbol{\theta}}(\mathbf{x})| = 2\left\|\mathrm{P}_\tau(\mathbf{x} \mid \tilde{\mathbf{x}}_0) - \mathrm{P}_{\boldsymbol{\theta}}(\mathbf{x})\right\|_{\mathrm{TV}},
\end{aligned} \tag{19}$$

where we have used the fact that $|\nabla_j \log \mathrm{P}_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}_\tau)| \le 1$ when $\boldsymbol{\psi}(\mathbf{x}) \in \{0,1\}^m$, for all $\mathbf{x} \in \{0,1\}^p$.

Therefore, by (19),

$$\begin{aligned}
\|\mathbb{E}_{\tilde{\mathbf{x}}_\tau}[\boldsymbol{\delta}(\boldsymbol{\theta}) \mid \tilde{\mathbf{x}}_0]\|_2 =& \sqrt{\sum_{j=1}^m |\mathbb{E}_{\tilde{\mathbf{x}}_\tau}[\delta_j(\boldsymbol{\theta}) \mid \tilde{\mathbf{x}}_0]|^2} \le \sqrt{m \times (2\left\|\mathrm{P}_\tau(\mathbf{x} \mid \mathbf{x}_0) - \mathrm{P}_{\boldsymbol{\theta}}(\mathbf{x})\right\|_{\mathrm{TV}})^2} \\
=& 2\sqrt{m}\left\|\mathrm{P}_\tau(\mathbf{x} \mid \mathbf{x}_0) - \mathrm{P}_{\boldsymbol{\theta}}(\mathbf{x})\right\|_{\mathrm{TV}}.
\end{aligned}$$

A.3.2 Proof of Lemma 2

Let $j \neq i$ be given. With $\xi_{ij} = \theta_{\min\{i,j\},\max\{i,j\}}$, consider

$$\begin{aligned}
\mathrm{P}_{\boldsymbol{\theta}}(X_i = 1 \mid \mathbf{X}_{-i}) =& \frac{\mathrm{P}_{\boldsymbol{\theta}}(X_i = 1, \mathbf{X}_{-i})}{\mathrm{P}_{\boldsymbol{\theta}}(X_i = 0, \mathbf{X}_{-i}) + \mathrm{P}_{\boldsymbol{\theta}}(X_i = 1, \mathbf{X}_{-i})} \\
=& \frac{1}{1 + \exp\left(-\theta_{ii} - \sum_{k\neq i} \xi_{i,k} X_k\right)} \\
=& \frac{1}{1 + \exp\left(-\theta_{ii} - \sum_{k\neq i, k\neq j} \xi_{i,k} X_k\right)\exp\left(-\xi_{i,j} X_j\right)} \\
=& g\left(\exp\left(-\xi_{i,j} X_j\right), b_1\right),
\end{aligned}$$

where

$$b = \exp\left(-\theta_{ii} - \sum_{k\neq i, k\neq j} \xi_{i,k} X_k\right) \in [r, s],$$

with

$$r = \exp\left(-\theta_{ii} - \sum_{k\neq i, k\neq j} \xi_{i,k} \max\left\{\mathrm{sgn}(\xi_{i,k}), 0\right\}\right), \quad s = \exp\left(-\theta_{ii} - \sum_{k\neq i, k\neq j} \xi_{i,k} \max\left\{-\mathrm{sgn}(\xi_{i,k}), 0\right\}\right).$$

Therefore,

$$\begin{aligned}
C_{ij} &= \max_{\mathbf{X},\mathbf{Y}\in N_j} \frac{1}{2}|\mathrm{P}_{\boldsymbol{\theta}}(X_i = 1 \mid \mathbf{X}_{-i}) - \mathrm{P}_{\boldsymbol{\theta}}(Y_i = 1 \mid \mathbf{Y}_{-i})| + \frac{1}{2}|\mathrm{P}_{\boldsymbol{\theta}}(X_i = 0 \mid \mathbf{X}_{-i}) - \mathrm{P}_{\boldsymbol{\theta}}(Y_i = 0 \mid \mathbf{Y}_{-i})| \\
&= \max_{\mathbf{X},\mathbf{Y}\in N_j} |\mathrm{P}_{\boldsymbol{\theta}}(X_i = 1 \mid \mathbf{X}_{-i}) - \mathrm{P}_{\boldsymbol{\theta}}(Y_i = 1 \mid \mathbf{Y}_{-i})| \\
&= \max_{\mathbf{X},\mathbf{Y}\in N_j} |g\left(\exp\left(-\xi_{i,j} X_j\right), b\right) - g\left(\exp\left(-\xi_{i,j} Y_j\right), b\right)| \\
&= \max_{\mathbf{X},\mathbf{Y}\in N_j} \frac{|\exp\left(-\xi_{i,j} X_j\right) - \exp\left(-\xi_{i,j} Y_j\right)| b}{\left(1 + b\exp\left(-\xi_{i,j} X_j\right)\right)\left(1 + b_1 \exp\left(-\xi_{i,j} Y_j\right)\right)} \\
&= \max_{\mathbf{X},\mathbf{Y}\in N_j} \frac{|\exp\left(-\xi_{i,j}\right) - 1| b}{\left(1 + b\exp\left(-\xi_{i,j}\right)\right)(1 + b)}.
\end{aligned}$$

Then following the Lemma 15 in Mitliagkas and Mackey [2017], we have

$$C_{ij} \leq \frac{|\exp\left(-\xi_{i,j}\right) - 1| b^*}{\left(1 + b_1^* \exp\left(-\xi_{i,j}\right)\right)(1 + b^*)}, \tag{20}$$

with $b^* = \max\left\{r, \min\left\{s, \exp\left(\frac{\xi_{i,j}}{2}\right)\right\}\right\}$.

A.4 Proof of Theorem 4

We are interested in concentrating $\|\boldsymbol{\delta}(\boldsymbol{\theta})\|_2$ around $\|\mathbb{E}_{\tilde{\mathbf{x}}_\tau}[\boldsymbol{\delta}(\boldsymbol{\theta}) \mid \tilde{\mathbf{x}}_0]\|_2$. To this end, we first consider concentrating $\delta_j(\boldsymbol{\theta})$ around $\mathbb{E}_{\tilde{\mathbf{x}}_\tau}[\delta_j(\boldsymbol{\theta}) \mid \tilde{\mathbf{x}}_0]$, where $j \in \{1, 2, \cdots, m\}$. Let q defined in Algorithm 2 be given. Then q trials of Gibbs sampling are run, resulting in $\left\{\delta_j^{(1)}(\boldsymbol{\theta}), \delta_j^{(2)}(\boldsymbol{\theta}), \cdots, \delta_j^{(q)}(\boldsymbol{\theta})\right\}$, and $\left\{\psi_j^{(1)}(\boldsymbol{\theta}), \psi_j^{(2)}(\boldsymbol{\theta}), \cdots, \psi_j^{(q)}(\boldsymbol{\theta})\right\}$ defined in Section 4.2, one element for each of the q trials. Since all the trials are independent, $\delta_j^{(i)}(\boldsymbol{\theta})$'s can be considered as i.i.d. samples with mean $\mathbb{E}_{\tilde{\mathbf{x}}_\tau}[\delta_j(\boldsymbol{\theta}) \mid \tilde{\mathbf{x}}_0]$. Furthermore, $\delta_j^{(i)}(\boldsymbol{\theta}) = \nabla_j \log \mathrm{P}_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}_\tau) \in [-1, 1]$ when $\boldsymbol{\psi}(\mathbf{x}) \in \{0, 1\}^m$, for all $\mathbf{x} \in \{0, 1\}^p$. Let $\beta_j > 0$ be given; we define the adversarial event:

$$E_j^q(\epsilon_j) = \left|\frac{1}{q}\sum_{i=1}^{q} \delta_j^{(i)}(\boldsymbol{\theta}) - \mathbb{E}_{\tilde{\mathbf{x}}_\tau}[\delta_j(\boldsymbol{\theta}) \mid \tilde{\mathbf{x}}_0]\right| > \epsilon_j, \tag{21}$$

with $j \in \{1, 2, \cdots, m\}$.

Define another random variable $Z_j = \frac{1+\delta_j(\boldsymbol{\theta})}{2}$ with samples $Z_j^{(i)} = \frac{1+\delta_j^{(i)}(\boldsymbol{\theta})}{2}$ and the sample variance $V_{Z_j} = \frac{V_{\delta_j}}{4} = \frac{V_{\psi_j}}{4}$.

Considering $Z \in [0, 1]$, we can apply Theorem 4 in Maurer and Pontil [2009] and achieve

$$\mathrm{P}\left(\left|\frac{1}{q}\sum_{i=1}^{q} Z_j^{(i)} - \mathbb{E}_{\tilde{\mathbf{x}}_\tau}[Z_j \mid \tilde{\mathbf{x}}_0]\right| > \frac{\epsilon_j}{2}\right) \leq 2\beta_j,$$

where

$$\frac{\epsilon_j}{2} = \sqrt{\frac{2V_{Z_j} \ln 2/\beta_j}{q}} + \frac{7\ln 2/\beta_j}{3(p-1)} = \sqrt{\frac{V_{\psi_j} \ln 2/\beta_j}{2q}} + \frac{7\ln 2/\beta_j}{3(p-1)}.$$

That is to say

$$\mathrm{P}\left(E_j^q(\epsilon_j)\right) \le 2\beta_j.$$

Now, for all $j \in \{1, 2, \cdots, m\}$, we would like $\frac{1}{m}\sum_{i=1}^{m} \delta_j^{(i)}(\boldsymbol{\theta})$ to be close to $\mathbb{E}_{\tilde{\mathbf{x}}_\tau}[\delta_j(\boldsymbol{\theta}) \mid \tilde{\mathbf{x}}_0]$. i.e.,

$$\left| \frac{1}{q}\sum_{i=1}^{q} \delta_j^{(i)}(\boldsymbol{\theta}) - \mathbb{E}_{\tilde{\mathbf{x}}_\tau}[\delta_j(\boldsymbol{\theta}) \mid \tilde{\mathbf{x}}_0] \right| \le \epsilon_j.$$

This concentrated event will occur with probability:

$$1 - \mathrm{P}\left(E_j(\epsilon_j)\right) \ge 1 - \mathrm{P}\left(E_j^q(\epsilon_j)\right) \ge 1 - 2\beta_j.$$

When all the concentrated events occur for each j,

$$\begin{aligned}
\|\boldsymbol{\delta}(\boldsymbol{\theta})\|_2 - \|\mathbb{E}_{\tilde{\mathbf{x}}_\tau}[\boldsymbol{\delta}(\boldsymbol{\theta}) \mid \tilde{\mathbf{x}}_0]\|_2 \le \|\boldsymbol{\delta}(\boldsymbol{\theta}) - \mathbb{E}_{\tilde{\mathbf{x}}_\tau}[\boldsymbol{\delta}(\boldsymbol{\theta}) \mid \tilde{\mathbf{x}}_0]\|_2 &= \left\| \frac{1}{q}\sum_{i=1}^{q} \boldsymbol{\delta}^{(i)}(\boldsymbol{\theta}) - \mathbb{E}_{\tilde{\mathbf{x}}_\tau}[\boldsymbol{\delta}(\boldsymbol{\theta}) \mid \tilde{\mathbf{x}}_0] \right\|_2 \\
&= \sqrt{\sum_{j=1}^{m} \left(\frac{1}{q}\sum_{i=1}^{q} \delta_j^{(i)}(\boldsymbol{\theta}) - \mathbb{E}_{\tilde{\mathbf{x}}_\tau}[\delta_j(\boldsymbol{\theta}) \mid \tilde{\mathbf{x}}_0] \right)^2} \le \sqrt{\sum_{j=1}^{m} \epsilon_j^2}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\|\boldsymbol{\delta}(\boldsymbol{\theta})\|_2 \le \|\mathbb{E}_{\tilde{\mathbf{x}}_\tau}[\boldsymbol{\delta}(\boldsymbol{\theta}) \mid \tilde{\mathbf{x}}_0]\|_2 + \sqrt{\sum_{j=1}^{m} \epsilon_j^2} &\le 2\sqrt{m}\,\|\mathrm{P}_\tau(\mathbf{x} \mid \tilde{\mathbf{x}}_0) - \mathrm{P}_{\boldsymbol{\theta}}(\mathbf{x})\|_{\mathrm{TV}} + \sqrt{\sum_{j=1}^{m} \epsilon_j^2} \\
&\le 2\sqrt{m}\left(\mathscr{G}(\mathbf{B}^\tau) + \sqrt{\frac{\sum_{j=1}^{m} \epsilon_j^2}{4m}} \right).
\end{aligned}$$

That is to say, we can conclude that (13) holds provided that all the concentrated events occur. Thus, the probability that (13) holds follows the inequality below:

$$\mathrm{P}\left(\|\boldsymbol{\delta}(\boldsymbol{\theta})\|_2 \le 2\sqrt{m}\left(\mathscr{G}(\mathbf{B}^\tau) + \sqrt{\frac{\sum_{j=1}^{m} \epsilon_j^2}{4m}} \right) \right) \ge 1 - \mathrm{P}\left(\bigcup_{j=1}^{m} E_j(\epsilon_j) \right) \ge 1 - \sum_{j=1}^{m} \mathrm{P}\left(E_j^q(\epsilon_j)\right) \ge 1 - 2\sum_{j=1}^{m} \beta_j.$$

A.5 Proof of Theorem 5

We consider the probability that the achieved objective function value decreases in the k^{th} iteration provided that the criterion TAY-CRITERION is satisfied:

$$\mathrm{P}\left(g(\boldsymbol{\theta}^{(k+1)}) < g(\boldsymbol{\theta}^{(k)}) \mid 2\sqrt{m}\mathscr{G}(\mathbf{B}^\tau) < \frac{1}{2}\|\boldsymbol{G}_\alpha(\boldsymbol{\theta}^{(k)})\|_2 \right).$$

Since $\|\boldsymbol{\delta}(\boldsymbol{\theta}^{(k)})\|_2 \le \frac{1}{2}\|\boldsymbol{G}_\alpha(\boldsymbol{\theta}^{(k)})\|_2$ provided in Theorem 1 is a sufficient condition for $g(\boldsymbol{\theta}^{(k+1)}) \le g(\boldsymbol{\theta}^{(k)})$, we have:

$$\begin{aligned}
&\mathrm{P}\left(g(\boldsymbol{\theta}^{(k+1)}) \le g(\boldsymbol{\theta}^{(k)}) \mid 2\sqrt{m}\mathscr{G}(\mathbf{B}^\tau) < \frac{1}{2}\|\boldsymbol{G}_\alpha(\boldsymbol{\theta}^{(k)})\|_2 \right) \\
\ge& \mathrm{P}\left(\|\boldsymbol{\delta}(\boldsymbol{\theta}^{(k)})\|_2 \le \frac{1}{2}\|\boldsymbol{G}_\alpha(\boldsymbol{\theta}^{(k)})\|_2 \mid 2\sqrt{m}\mathscr{G}(\mathbf{B}^\tau) < \frac{1}{2}\|\boldsymbol{G}_\alpha(\boldsymbol{\theta}^{(k)})\|_2 \right) \\
=& 1 - \mathrm{P}\left(\|\boldsymbol{\delta}(\boldsymbol{\theta}^{(k)})\|_2 > \frac{1}{2}\|\boldsymbol{G}_\alpha(\boldsymbol{\theta}^{(k)})\|_2 \mid 2\sqrt{m}\mathscr{G}(\mathbf{B}^\tau) < \frac{1}{2}\|\boldsymbol{G}_\alpha(\boldsymbol{\theta}^{(k)})\|_2 \right) \\
\ge& 1 - \mathrm{P}\left(\|\boldsymbol{\delta}(\boldsymbol{\theta}^{(k)})\|_2 - \|\mathbb{E}_{\tilde{\mathbf{x}}_\tau}[\boldsymbol{\delta}(\boldsymbol{\theta}) \mid \tilde{\mathbf{x}}_0]\|_2 > \frac{1}{2}\|\boldsymbol{G}_\alpha(\boldsymbol{\theta}^{(k)})\|_2 - 2\sqrt{m}\mathscr{G}(\mathbf{B}^\tau) \mid 2\sqrt{m}\mathscr{G}(\mathbf{B}^\tau) < \frac{1}{2}\|\boldsymbol{G}_\alpha(\boldsymbol{\theta}^{(k)})\|_2 \right)
\end{aligned}$$

$$\geq 1-\sum_{j=1}^{m} \mathrm{P}\left(E_j^q(\frac{1}{2\sqrt{m}}\|\boldsymbol{G}_\alpha(\boldsymbol{\theta}^{(k)})\|_2 - 2\mathscr{G}(\mathbf{B}^\tau)) \mid 2\sqrt{m}\mathscr{G}(\mathbf{B}^\tau) < \frac{1}{2}\|\boldsymbol{G}_\alpha(\boldsymbol{\theta}^{(k)})\|_2\right),$$

where $E_j^q\left(\frac{1}{2\sqrt{m}}\|\boldsymbol{G}_\alpha(\boldsymbol{\theta}^{(k)})\|_2 - 2\mathscr{G}(\mathbf{B}^\tau)\right)$ is defined in (21) and in the 4^{th} line we apply (12). As q approaches infinity, by the weak law of large numbers, we have

$$\lim_{q\to\infty} \mathrm{P}\left(E_j^q\left((\frac{1}{2\sqrt{m}}\|\boldsymbol{G}_\alpha(\boldsymbol{\theta}^{(k)})\|_2 - 2\mathscr{G}(\mathbf{B}^\tau)\right)\right) = 0.$$

Then,

$$\begin{aligned}
&\lim_{q\to\infty} \mathrm{P}\left(g(\boldsymbol{\theta}^{(k+1)}) < g(\boldsymbol{\theta}^{(k)}) \mid 2\sqrt{m}\mathscr{G}(\mathbf{B}^\tau) < \frac{1}{2}\|\boldsymbol{G}_\alpha(\boldsymbol{\theta}^{(k)})\|_2\right) \\
\geq& 1-\lim_{q\to\infty}\sum_{j=1}^{m} \mathrm{P}\left(E_j^q(\frac{1}{2\sqrt{m}}\|\boldsymbol{G}_\alpha(\boldsymbol{\theta}^{(k)})\|_2 - 2\mathscr{G}(\mathbf{B}^\tau)) \mid 2\sqrt{m}\mathscr{G}(\mathbf{B}^\tau) < \frac{1}{2}\|\boldsymbol{G}_\alpha(\boldsymbol{\theta}^{(k)})\|_2\right) = 1.
\end{aligned}$$

A.6 Proof of Theorem 6

According to Theorem 2, we only need to show

$$\lim_{q\to\infty} \mathrm{P}\left(g(\boldsymbol{\theta}^{(k+1)}) \leq g(\boldsymbol{\theta}^{(k)})\right) = 1,$$

for $k = 1, 2, \cdots, \kappa - 1$.

By a union bound, the following inequality is true:

$$\lim_{q\to\infty} \mathrm{P}\left(g(\boldsymbol{\theta}^{(k+1)}) \leq g(\boldsymbol{\theta}^{(k)})\right) \leq 1 - \sum_{k=1}^{\kappa-1}\lim_{q\to\infty} \mathrm{P}\left(g(\boldsymbol{\theta}^{(k+1)}) > g(\boldsymbol{\theta}^{(k)})\right).$$

Notice that, following TAY, we always have:

$$\mathrm{P}\left(2\sqrt{m}\mathscr{G}(\mathbf{B}^\tau) < \frac{1}{2}\|\boldsymbol{G}_\alpha(\boldsymbol{\theta}^{(k)})\|_2\right) = 1,$$

suggesting

$$\lim_{q\to\infty} \mathrm{P}\left(g(\boldsymbol{\theta}^{(k+1)}) > g(\boldsymbol{\theta}^{(k)})\right) = \lim_{q\to\infty} \mathrm{P}\left(g(\boldsymbol{\theta}^{(k+1)}) > g(\boldsymbol{\theta}^{(k)}) \mid 2\sqrt{m}\mathscr{G}(\mathbf{B}^\tau) < \frac{1}{2}\|\boldsymbol{G}_\alpha(\boldsymbol{\theta}^{(k)})\|_2\right) = 0,$$

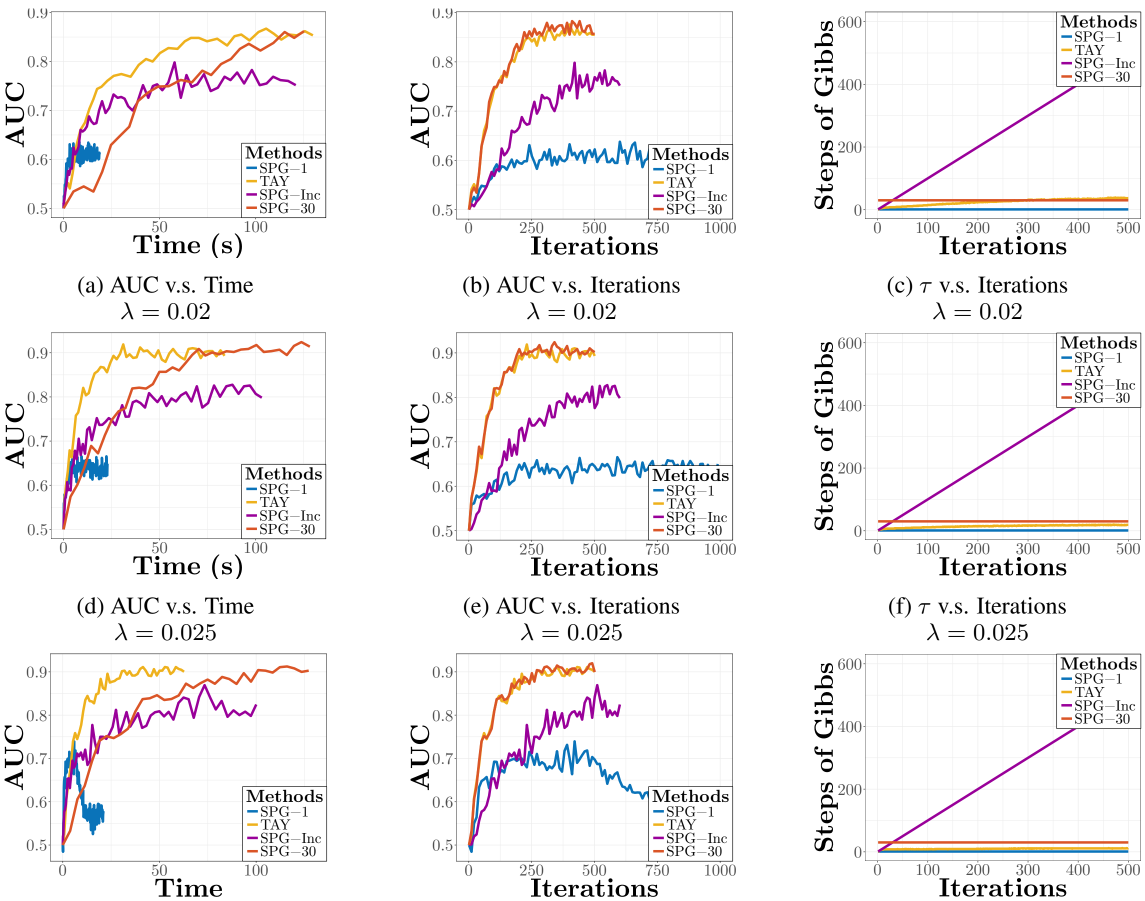
where the equality is due to Theorem 5.

Finally, with Theorem 2, we can finish the proof.

B Experiments

B.1 Comparison with SPG-based Methods

In this section, we consider the effect of the regularization parameter λ. Specifically, we apply the methods mentioned the Section 7.1 with different λs. The results are reported in Figure 4 and Figure 5.



(a) AUC v.s. Time $\lambda = 0.02$

(b) AUC v.s. Iterations $\lambda = 0.02$

(c) τ v.s. Iterations $\lambda = 0.02$

(d) AUC v.s. Time $\lambda = 0.025$

(e) AUC v.s. Iterations $\lambda = 0.025$

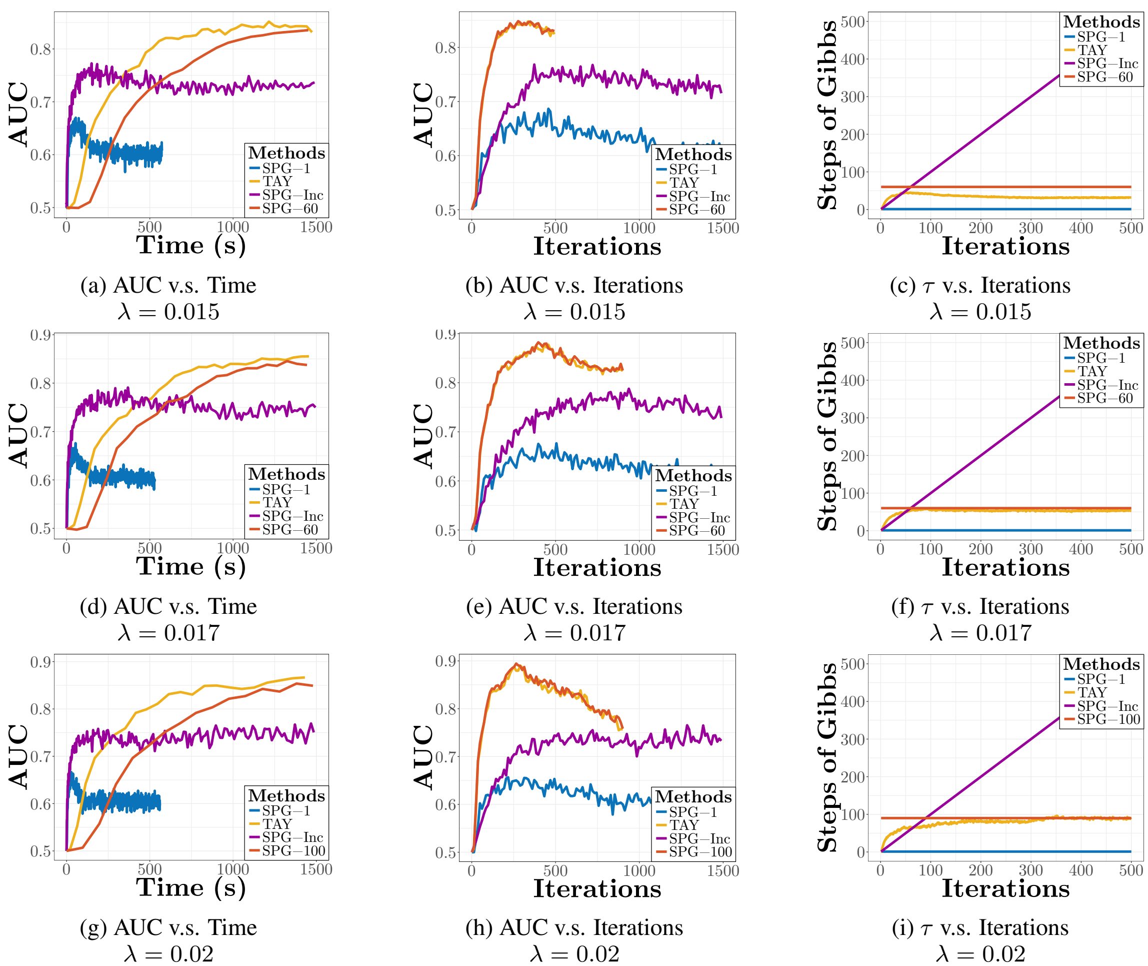
(f) τ v.s. Iterations $\lambda = 0.025$

(g) AUC v.s. Time $\lambda = 0.03$

(h) AUC v.s. Iterations $\lambda = 0.03$

(i) τ v.s. Iterations $\lambda = 0.03$

Figure 4: Area under curve (AUC) and the steps of Gibbs sampling (τ) for the structure learning of a 10-node network with different λ's.



(a) AUC v.s. Time $\lambda = 0.015$

(b) AUC v.s. Iterations $\lambda = 0.015$

(c) τ v.s. Iterations $\lambda = 0.015$

(d) AUC v.s. Time $\lambda = 0.017$

(e) AUC v.s. Iterations $\lambda = 0.017$

(f) τ v.s. Iterations $\lambda = 0.017$

(g) AUC v.s. Time $\lambda = 0.02$

(h) AUC v.s. Iterations $\lambda = 0.02$

(i) τ v.s. Iterations $\lambda = 0.02$

Figure 5: Area under curve (AUC) and the steps of Gibbs sampling (τ) for the structure learning of a 20-node network with different λ's.

B.2 Comparison with the Pseudo-likelihood Method

We compare TYA with the pseudo-likelihood method (Pseudo) under the same parameter configuration introduced in Section 7.1. Note that the two methods achieve a comparable performance: Pseudo is slightly better with 10 nodes and TAY outperforms a little with 20 nodes. This is consistent with the theoretical result that the two inductive principles are both sparsistent.

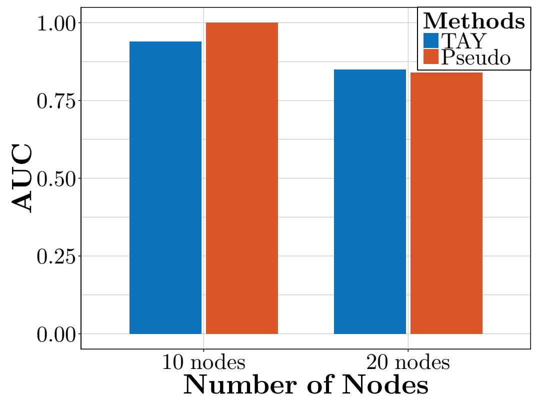


Figure 6: Area under curve (AUC) and for the structure learning of a 20-node network.