Supplementary Material for "On Estimation in Latent Variable Models"

Guanhua Fang, Ping Li
Cognitive Computing Lab
Baidu Research
10900 NE 8th St Bellevue WA 98004 USA
{guanhuafang, liping11}@baidu.com

Summary In this supplementary file, we collect the technical proofs for results stated in the main paper. Throughout the sequel, we will adopt the following notations. We let θ denote the generic model parameter. We also let Y be a random variable representing the observed data and Z be a random variable representing the latent unobserved variable. We use y and z to denote their realizations, respectively. Subscript i is used to indicate the i-th individual. We use $\|x\|$ and $\|x\|_1$ to represent ℓ_2- and ℓ_1-norm of vector x. For random sequences a_n and b_n, $a_n = O_p(b_n)$ represents that a_n is stochastically bounded by Kb_n for a sufficiently large constant K; $a_n = o_p(b_n)$ represents a_n/b_n converges to 0 with probability tending to 1. Moreover, $a = O(b)$ means there exists a constant K such that $a \leq Kb$; $a = \Omega(b)$ means there exists a sufficiently large constant K such that $a \geq Kb$; $a \gg b$ means that there exists a sufficiently K such that $a \geq Kb$. We use ∇f ($\nabla^2 f, \nabla^3 f$) to represent the first (second, third) derivative of function f with respect to θ. Lastly, constants c, C may be different from the place to place.

1. Proof of Theorem 1

We first define the following additional notations.

- Individualized gradient: $\nabla f_i(\theta) = -\nabla \log L_i(\theta)$, full gradient: $\nabla F_n(\theta) = \frac{1}{n}\sum_i \nabla f_i(\theta)$. (We may write $\nabla F_n(\theta) = \nabla F(\theta)$ for simplicity.)

- Individualized stochastic gradient: $\nabla H_i(\theta, z_i) = -\nabla \log p_\theta(y_i|z_i)$, batch stochastic gradient $\nabla H_B(\theta) = \frac{1}{n}\sum_{i\in B} \nabla H_i(\theta, z_i)$.

- We further write $\nabla f_i(\theta, \theta') = \mathbb{E}_{z_i \sim p_{\theta'}(z|y_i)} \nabla H_i(\theta, z_i)$ and $\nabla F_n(\theta, \theta') = \frac{1}{n}\sum_i \nabla f_i(\theta, \theta')$. Then it is easy to see that $\nabla f_i(\theta, \theta') = \nabla f_i(\theta)$ and $\nabla F_n(\theta, \theta') = \nabla F_n(\theta)$.

Bound of v_t^{s+1}: We first consider to give the upper bound of $\mathbb{E}\|v_t^{s+1}\|$. (Here expectation $\mathbb{E}$ is the conditional expectation which is only taken over all i_t^s's and $z_{i_t^{s+1}}$'s given other variables.) For fixed iteration s and update index t, we further define $\zeta_t^{s+1} = H(\theta_t^{s+1}, z_{i_t^{s+1}}) - H(\tilde\theta^s, z_{i_t^s})$. Then, by the definition of v_t^{s+1}, we have $v_t^{s+1} = \xi_t^{s+1} + \tilde\nabla f^{s+1} = \xi_t^{s+1} + \nabla H_{B^s}(\tilde\theta^s)$ according to the definition of our new notation. By taking expectation with respect to i_t^{s+1} and $z_{i_t^{s+1}}$, we have

$$\mathbb{E}_{i_t^{s+1}, z_{i_t^{s+1}}} v_t^{s+1} = \nabla F_n(\theta_t^{s+1}) - \nabla F_n(\tilde\theta^s, \theta_t^{s+1}) + \nabla H_{B^s}(\tilde\theta^s) := H_t^{s+1}. \tag{8}$$

Thus, we can compute

$$\begin{aligned}
\mathbb{E}_{i_t^{s+1}, z_{i_t^{s+1}}}[\|v_t^{s+1}\|^2] &= \mathbb{E}_{i_t^{s+1}, z_{i_t^{s+1}}}[\|\zeta_t^{s+1} + \tilde{\nabla} f^{s+1}\|^2] \\
&= \mathbb{E}_{i_t^{s+1}, z_{i_t^{s+1}}}[\|\zeta_t^{s+1} + \tilde{\nabla} f^{s+1} - H_t^{s+1} + H_t^{s+1}\|^2] \\
&\leq 2\mathbb{E}_{i_t^{s+1}, z_{i_t^{s+1}}}[\|H_t^{s+1}\|^2] + 2\mathbb{E}_{i_t^{s+1}, z_{i_t^{s+1}}}[\|\zeta_t^{s+1} - \mathbb{E}_{i_t^{s+1}, z_{i_t^{s+1}}}[\zeta_t^{s+1}]\|^2] \\
&\leq 2\|H_t^{s+1}\|^2 + 2\mathbb{E}_{i_t^{s+1}, z_{i_t^{s+1}}}[\|\zeta_t^{s+1}\|^2] \\
&\leq 2[\|H_t^{s+1}\|^2] + 2L^2\|\theta_t^{s+1} - \tilde{\theta}^s\|^2 \qquad (9) \\
&\leq 4[\|\nabla F_n(\theta_t^{s+1})\|^2 + \|\eta\|^2] + 2L^2[\|\theta_t^{s+1} - \tilde{\theta}^s\|^2] \qquad (10) \\
&\leq 2C\|\nabla F_n(\theta_t^{s+1})\|^2 + 2L^2\|\theta_t^{s+1} - \tilde{\theta}^s\|^2, \qquad (11)
\end{aligned}$$

by adjusting constant C and using the fact that $\mathbb{E}[\|\nabla F(\theta_t^{s+1})\|^2] = \Omega(1/n_1 + (m\gamma)^2)$ before the termination of the algorithm and $\|\eta\|^2$ is $O(1/n_1 + (m\gamma)^2)$ (which will be shown in the next paragraph). Here (9) uses the fact that the density function is smooth and hence is L-lipschitz continuous for some positive L. Inequality (10) holds due to the fact that $\|a+b\|^2 \leq \|a\|^2 + \|b\|^2$, where we write $\eta = \nabla F_n(\theta_t^{s+1}) - H_t^{s+1}$. Therefore, we obtain

$$\mathbb{E}[\|v_t^{s+1}\|^2] = 2C\mathbb{E}\|\nabla F_n(\theta_t^{s+1})\|^2 + 2L^2\mathbb{E}\|\theta_t^{s+1} - \tilde{\theta}^s\|^2. \qquad (12)$$

Difference between $\nabla F_n(\theta_t^{s+1})$ and H_t^{s+1}: By straightforward calculation, we can find that

$$\begin{aligned}
\|\nabla F_n(\theta_t^{s+1}) - H_t^{s+1}\| &= \|\nabla F_n(\tilde{\theta}^s, \theta_t^{s+1}) - \nabla H_{B^s}(\tilde{\theta}^s)\| \\
&= \|\nabla F_n(\tilde{\theta}^s, \theta_t^{s+1}) - \nabla F_n(\tilde{\theta}^s, \tilde{\theta}^s) + \nabla F_n(\tilde{\theta}^s) - \nabla H_{B^s}(\tilde{\theta}^s)\| \\
&= \|\nabla F_n(\tilde{\theta}^s, \theta_t^{s+1}) - \nabla F_n(\tilde{\theta}^s, \tilde{\theta}^s)\| + \|\nabla F_n(\tilde{\theta}^s) - \nabla H_{B^s}(\tilde{\theta}^s)\| \\
&\leq C\|\tilde{\theta}^s - \theta_t^{s+1}\| + \|\nabla F_n(\tilde{\theta}^s) - \nabla H_{B^s}(\tilde{\theta}^s)\|. \qquad (13)
\end{aligned}$$

Note that $\mathbb{E}\nabla F_n(\tilde{\theta}^s) = \mathbb{E}\nabla H_{B^s}(\tilde{\theta}^s) = \mathbb{E}_y \nabla \log L(\theta)$. Therefore, by Hoeffding's concentration inequality, we have that

$$P(\|\nabla F_n(\tilde{\theta}^s) - \mathbb{E}_y \nabla \log L(\theta)\| \geq \frac{C_1}{\sqrt{n}}) \leq \exp\{-2C_1^2/m_1\}$$

and

$$P(\|\nabla H_{B^s}(\tilde{\theta}^s) - \mathbb{E}_y \nabla \log L(\theta)\| \geq \frac{C_2}{\sqrt{n_1}}) \leq \exp\{-2C_2^2/m_2\}$$

where m_1 and m_2 are the upper bounds for $|\nabla \log L(\theta)|$ and $|\nabla \log p_\theta(y|z)|$. Such constants exist by the compactness assumption. Therefore, with high probability, we have that

$$\begin{aligned}
\|\eta\| = \|\nabla F_n(\theta_t^{s+1}) - H_t^{s+1}\| &\leq C\|\tilde{\theta}^s - \theta_t^{s+1}\| + \frac{C_1}{\sqrt{n}} + \frac{C_2}{\sqrt{n_1}} \\
&\leq C'(m\gamma + \frac{1}{\sqrt{n_1}}), \qquad (14)
\end{aligned}$$

where the last inequality uses the fact that $\|\tilde{\theta}^s - \theta_t^{s+1}\|$ is at most of order $m\gamma$.

By this, we can further obtain that

$$\begin{aligned}
\langle \nabla F_n(\theta_t^{s+1}), H_t^{s+1} \rangle &= \langle \nabla F_n(\theta_t^{s+1}), \nabla F_n(\theta_t^{s+1}) \rangle - \langle \nabla F_n(\theta_t^{s+1}), \nabla F_n(\theta_t^{s+1}) - H_t^{s+1} \rangle \\
&\geq \|\nabla f(\theta_t^{s+1})\|^2 - \|\nabla F_n(\theta_t^{s+1})\|\|\nabla F_n(\theta_t^{s+1}) - H_t^{s+1}\| \\
&\geq c\|\nabla F_n(\theta_t^{s+1})\|^2 \qquad (15)
\end{aligned}$$

by adjusting constant c and using the fact that $\mathbb{E}[\|\nabla F_n(\theta_t^{s+1})\|^2] = \Omega(\|\eta\|)$ before the termination of the algorithm.

Descent Inequality: By smoothness of $F(\theta)$, we then have

$$\begin{aligned}
\mathbb{E}[F(\theta_{t+1}^{s+1})] &\leq \mathbb{E}[F(\theta_t^{s+1}) + \langle \nabla F(\theta_t^{s+1}), \theta_{t+1}^{s+1} - \theta_t^{s+1}\rangle + \frac{L}{2}\|\theta_{t+1}^{s+1} - \theta_t^{s+1}\|^2] \\
&= \mathbb{E}[F(\theta_t^{s+1}) - \gamma\langle \nabla F(\theta_t^{s+1}), v_t^{s+1}\rangle + \frac{L\gamma^2}{2}\|v_t^{s+1}\|^2] \\
&= \mathbb{E}[F(\theta_t^{s+1}) - \gamma\langle \nabla F(\theta_t^{s+1}), H_t^{s+1}\rangle + \frac{L\gamma^2}{2}\|v_t^{s+1}\|^2], \qquad (16)
\end{aligned}$$

for some constant L.

Consider the following Lyapunov function (Reddi et al., 2016)

$$R_t^{s+1} := \mathbb{E}[F(\theta_t^{s+1}) + c_t\|\theta_t^{s+1} - \tilde{\theta}^s\|^2],$$

where c_t is defined recursively in (19). We can compute that

$$\begin{aligned}
& \mathbb{E}[\|\theta_{t+1}^{s+1} - \tilde{\theta}^s\|^2] \\
= \; & \mathbb{E}[\|\theta_{t+1}^{s+1} - \theta_t^{s+1}\|^2 + \|\theta_t^{s+1} - \tilde{\theta}^s\|^2 + 2\langle \theta_{t+1}^{s+1} - \theta_t^{s+1}, \theta_t^{s+1} - \tilde{\theta}^s\rangle] \\
= \; & \mathbb{E}[\gamma^2\|v_t^{s+1}\|^2 + \|\theta_t^{s+1} - \tilde{\theta}^s\|^2] + 2\gamma\mathbb{E}[\langle H_t^{s+1}, \theta_t^{s+1} - \tilde{\theta}^s\rangle] \\
\leq \; & \mathbb{E}[\gamma^2\|v_t^{s+1}\|^2 + \|\theta_t^{s+1} - \tilde{\theta}^s\|^2] + 2\gamma\mathbb{E}[\frac{1}{2\beta_t}\|H_t^{s+1}\|^2 + \frac{\beta_t}{2}\|\theta_t^{s+1} - \tilde{\theta}^s\|^2] \\
\leq \; & \mathbb{E}[\gamma^2\|v_t^{s+1}\|^2 + \|\theta_t^{s+1} - \tilde{\theta}^s\|^2] + 2\gamma\mathbb{E}[\frac{c_2}{2\beta_t}\|\nabla f(\theta_t^{s+1})\|^2 + \frac{\beta_t}{2}\|\theta_t^{s+1} - \tilde{\theta}^s\|^2], \qquad (17)
\end{aligned}$$

where β_t will be determined later. Combining (16) and (17), we then have

$$\begin{aligned}
R_{t+1}^{s+1} &= \mathbb{E}[F(\theta_{t+1}^{s+1}) + c_{t+1}\|\theta_{t+1}^{s+1} - \tilde{\theta}^s\|^2] \\
&\leq \mathbb{E}[F(\theta_t^{s+1}) - \gamma\langle \nabla F(\theta_t^{s+1}), H_t^{s+1}\rangle + \frac{L\gamma^2}{2}\|v_t^{s+1}\|^2] \\
&\quad + c_{t+1}(\mathbb{E}[\gamma^2\|v_t^{s+1}\|^2 + \|\theta_t^{s+1} - \tilde{\theta}^s\|^2] + 2\gamma\mathbb{E}[\frac{1}{2\beta_t}\|H_t^{s+1}\|^2 + \frac{\beta_t}{2}\|\theta_t^{s+1} - \tilde{\theta}^s\|^2]) \\
&= \mathbb{E}[F(\theta_t^{s+1}) - (c\gamma - c_2\frac{c_{t+1}\gamma}{\beta_t})\|\nabla F(\theta_t^{s+1})\|^2 \\
&\quad + (\frac{L\gamma^2}{2} + c_{t+1}\gamma^2)\mathbb{E}[\|v_t^{s+1}\|^2] + (c_{t+1} + c_{t+1}\gamma\beta_t)\mathbb{E}[\|\theta_t^{s+1} - \tilde{\theta}^s\|^2]. \qquad (18)
\end{aligned}$$

Together with the bound on $\mathbb{E}\|v_t^{s+1}\|^2$, we then have

$$\begin{aligned}
R_{t+1}^{s+1} &\leq \mathbb{E}[F(\theta_t^{s+1})] \\
&\quad - (c\gamma - \frac{c_2 c_{t+1}\gamma}{\beta_t} - C\gamma^2 L - 2Cc_{t+1}\gamma^2)\mathbb{E}[\|\nabla F(\theta_t^{s+1})\|^2] \\
&\quad + (c_{t+1}(1 + \gamma\beta_t + 2\gamma^2 L^2) + \gamma^2 L^3)E[\|\theta_t^{s+1} - \tilde{\theta}^s\|^2] \\
&= R_t^{s+1} - (c\gamma - \frac{c_2 c_{t+1}\gamma}{\beta_t} - C\gamma^2 L - 2Cc_{t+1}\gamma^2)\mathbb{E}[\|\nabla F(\theta_t^{s+1})\|^2],
\end{aligned}$$

where we define the recursive relationship between c_t's, i.e.,

$$c_t = c_{t+1}(1 + \gamma\beta_t + 2\gamma^2 L^2) + \gamma^2 L^3. \qquad (19)$$

For notational simplicity, we define

$$\Gamma_t = c\gamma - \frac{c_2 c_{t+1}\gamma}{\beta_t} - C\gamma^2 L - 2Cc_{t+1}\gamma^2$$

and $\gamma_{min} = \min_t \Gamma_t$. We add up (19) over t from 0 to $t-1$ and get

$$\gamma_{min} \sum_{t=0}^{m-1} \mathbb{E}[\|\nabla F(\theta_t^{s+1})\|^2] \leq R_0^{s+1} - R_m^{s+1}.$$

Note that $c_m = 0$, then $R_m^{s+1} = \mathbb{E}[F(\theta_m^{s+1})] = \mathbb{E}[F(\tilde{\theta}^{s+1})]$ and that $R_0^{s+1} = \mathbb{E}[F(\theta_0^{s+1})] = \mathbb{E}[F(\tilde{\theta}^s)]$. Therefore, we have

$$\frac{1}{T} \sum_{s=0}^{S-1} \sum_{t=0}^{m-1} \mathbb{E}[\|\nabla F(\theta_t^{s+1})\|^2] \leq (F(\theta^0) - F(\bar{\theta}))/(T\gamma_{min}). \tag{20}$$

Recall that $T_{min}(\epsilon)$ is $\arg\min_t \min_s\{\mathbb{E}\|\nabla F(\theta_t^s)\|^2 \leq \epsilon\}$. Then (20) gives us that $T_{min}(\epsilon) \leq \frac{F(\theta^0)-F(\bar{\theta})}{m\gamma_{min}\epsilon}$ with high probability. This completes the proof of Theorem 1.

Choice of n_1, m and γ: We take β_t as the constant β (i.e., free of t and s) and let $r = 2\gamma^2 L^2 + \gamma\beta$, $\gamma = \frac{1}{Ln^\alpha}$, $m = n^{\alpha_1}$, $n_1 = n^{2(\alpha-\alpha_1)}$, $\beta = Ln^{-\alpha/2}$. Then r is bounded by $\gamma\beta + L^2\gamma^2 = O(\gamma\beta)$. We can compute c_0 which is bounded by

$$\begin{aligned} c_0 &= L^3\gamma^2 \frac{(1+r)^m - 1}{r} \\ &\leq L^3\gamma^2 \frac{(1+\gamma\beta)^m - 1}{\gamma\beta} \\ &= Ln^{-\alpha/2}((1+\gamma\beta)^m - 1) \\ &\leq \mu L n^{-\alpha/2}, \end{aligned} \tag{21}$$

where $\mu = O(\gamma\beta m)$ which goes to 0 as $n \to \infty$. By the definition of γ_{min}, we can compute

$$\begin{aligned} \gamma_{min} &= \min_t\{c\gamma - \frac{c_2 c_{t+1}\gamma}{\beta_t} - C\gamma^2 L - 2Cc_{t+1}\gamma^2\} \\ &\geq c\gamma - \frac{c_0\gamma}{\beta} - \gamma^2 L - 2c_0\gamma^2 \\ &\geq \frac{c'}{Ln^\alpha} \end{aligned} \tag{22}$$

holds for some constant c'. Here the last inequality holds since that c_0/β is upper bounded by some constant times $(1+\gamma\beta)^m - 1$ which is $o(1)$, $\gamma^2 L \ll \gamma$ and $\gamma^2 \ll \gamma$.

Therefore, it gives $T(\epsilon) \leq C\frac{n^\alpha(F(\theta^0)-F(\bar{\theta}))}{m\epsilon}$. This concludes the proof of Theorem 1. Taking $n_1 = n^{\alpha_1}$, then the computational complexity will be $(n_1+m)\frac{n^\alpha}{m}$, which is n^α if $\alpha_1 \geq 2\alpha/3$ and $n^{2(\alpha-\alpha_1)}n^\alpha/n^{\alpha_1} = n^{3\alpha-3\alpha_1}$ if $\alpha_1 < 2\alpha/3$. Thus the total computational complexity is simplified as $C\frac{n^\alpha(F(\theta^0)-F(\bar{\theta}))}{\epsilon}$ by taking $m = n^{2\alpha/3}$ and $n_1 = n^{2\alpha/3}$. This gives Corollary 1.

2. Proof of Theorem 2

By Corollary 1, we know that $\|\nabla F(\hat{\theta})\|^2 = O_p(n^{-2\alpha/3})$. By Taylor expansion, we have that

$$\nabla F(\hat{\theta}) = \nabla F(\bar{\theta}) + \nabla^2 F(\check{\theta})(\hat{\theta} - \bar{\theta}) = \nabla^2 F(\check{\theta})(\hat{\theta} - \bar{\theta}), \tag{23}$$

where $\bar{\theta}$ is $\arg\min_\theta F(\theta)$ (also known as the maximal likelihood estimator) and $\check{\theta}$ is a point between $\hat{\theta}$ and $\bar{\theta}$. Since both $\hat{\theta}$ and $\check{\theta}$ are consistent estimator for θ^*, thus $\nabla^2 F(\check{\theta}) = I(\theta^*) + o_p(1)$ where $I(\theta^*)$ is the information matrix. Thus, $\|\hat{\theta} - \bar{\theta}\|^2 = O_p(n^{-2\alpha/3})$ as well.

Expand $\nabla F(\theta)$ at $\hat{\theta}$, we have

$$0 = \nabla F(\bar{\theta}) = \nabla F(\hat{\theta}) + \nabla^2 F(\hat{\theta})(\bar{\theta} - \hat{\theta}) + \frac{1}{2}\nabla^3 F(\xi)(\bar{\theta} - \hat{\theta})^2. \tag{24}$$

Since we have already know that $|\nabla H(\hat{\theta}) - \nabla f(\hat{\theta})| = O_p(\frac{1}{\sqrt{n}})$ and $|\nabla^2 H(\hat{\theta}) - \nabla^2 f(\hat{\theta})| = O_p(\frac{1}{\sqrt{n}})$. Plugging the formula of θ^{r_1} into (24), we get

$$\begin{aligned} 0 &= O_p(\frac{1}{\sqrt{n}}) + \nabla H(\hat{\theta}) + \nabla^2 F(\hat{\theta})(\bar{\theta} - \hat{\theta}) + \frac{1}{2}\nabla^3 F(\xi)(\bar{\theta} - \hat{\theta})^2 \\ &= O_p(\frac{1}{\sqrt{n}}) + \nabla^2 H(\hat{\theta})(\hat{\theta} - \theta^{r_1}) + \nabla^2 f(\hat{\theta})(\bar{\theta} - \hat{\theta}) + \frac{1}{2}\nabla^3 f(\xi)(\bar{\theta} - \hat{\theta})^2 \\ &= O_p(\frac{1}{\sqrt{n}}) + \nabla^2 H(\hat{\theta})(\bar{\theta} - \theta^{r_1}) + \frac{1}{2}\nabla^3 f(\xi)(\bar{\theta} - \hat{\theta})^2 \end{aligned} \tag{25}$$

Then we arrive at

$$\|\bar{\theta} - \theta^{r_1}\| = (\sigma_{min}(\nabla^2 H(\hat{\theta})))^{-1}\big(O_p(\frac{1}{\sqrt{n}}) + \frac{1}{2}|\nabla^3 f(\xi)|\|\bar{\theta} - \hat{\theta}\|^2\big).$$

We know that the algorithm returns $\hat{\theta}$ satisfy that $\|\hat{\theta} - \theta^*\| = O_p(\frac{1}{n^{1/3\alpha}})$. Therefore, we arrive at

$$\|\bar{\theta} - \theta^{r_1}\| = O_p(\frac{1}{\sqrt{n}} + n^{-2\alpha/3}).$$

Thus, when $3/4 < \alpha < 1$, we get $\|\bar{\theta} - \theta^{r_1}\| = O_p(\frac{1}{\sqrt{n}})$. It is known that MLE is root n-consistent. Thus we finally get

$$\|\theta^{r_1} - \theta^*\| = O_p(\frac{1}{\sqrt{n}}).$$

By two-step refinement, we recall the formula

$$\theta^{r_2} = \theta^{r_1} - \frac{\nabla H(\theta^{r_1})}{\nabla^2 H(\theta^{r_1})}. \tag{26}$$

Next we can show the normality of θ^{r_2}. By Taylor expansion, we know

$$\nabla H(\theta^{r_1}) = \nabla H(\theta^*) + (\theta^{r_1} - \theta^*)\nabla H^2(\theta^*) + \frac{1}{2}(\theta^{r_1} - \theta^*)\nabla^3 H(\xi), \tag{27}$$

where ξ lies between θ^* and θ^{r_1}. Put (26) into the above equation, we can get

$$\begin{aligned} \sqrt{n}(\theta^{r_2} - \theta^*) &= \frac{(1/\sqrt{n})\nabla H(\theta^*)}{-(1/n)\nabla^2 H(\theta^{r_1})} + \sqrt{n}(\theta^{r_1} - \theta^*) \\ &\quad \cdot\big[1 - \frac{\nabla^2 H(\theta^*)}{\nabla^2 H(\theta^{r_1})} - \frac{1}{2}(\theta^{r_1} - \theta^*)\frac{\nabla^3 H(\xi)}{\nabla^2 H(\theta^{r_1})}\big] \end{aligned} \tag{28}$$

after simplification. Then, we can see that the first term of (28) converges to $N(0, I^{-1}(\theta^*)V(\theta^*)I^{-1}(\theta^*))$. The second term of (28) is $o_p(1)$ since that $\sqrt{n}(\theta^{r_1} - \theta^*)$ is $O_p(1)$, $1 - \frac{\nabla^2 H(\theta^*)}{\nabla^2 H(\theta^{r_1})} = o_P(1)$ and $(\theta^{r_1} - \theta^*)\frac{\nabla^3 H(\xi)}{\nabla^2 H(\hat{\theta}^{r_1})}$ is $o_p(1)$. Lastly, by Slutsky Theorem, we get

$$\sqrt{n}(\theta^{r_2} - \theta^*) \to N(0, I^{-1}(\theta^*)V(\theta^*)I^{-1}(\theta^*)).$$

3. Proof of Theorem 3

We require the following lemmas for convergence analysis under non-smooth setting.

Lemma 1 *Let R be a closed convex function and $x, y \in dom(R)$. Then it holds*

$$\|prox_R(x) - prox_R(y)\| \leq \|x - y\|.$$

Lemma 2 *Let $P(\theta) = F(\theta) + R(\theta)$, where $\nabla F(\theta)$ is L-Lipschitz continuous, and $F(\theta)$ and $R(\theta)$ are strongly convex with parameter μ_F and μ_R. For any θ in domain and vector v, define*

$$\theta^+ = prox_{\gamma R}(\theta - \gamma v),\ g = \frac{1}{\gamma}(\theta - \theta^+),\ \Delta = v - \nabla F(\theta),$$

then it holds that

$$P(y) \geq P(\theta^+) + g^T(y-\theta) + \frac{\eta}{2}\|g\|^2 + \frac{\mu_F}{2}\|y-\theta\|^2 + \frac{\mu_R}{2}\|y - \theta^+\|^2 + \Delta^T(\theta^+ - y) \tag{29}$$

for any y in the domain and $0 < \gamma < 1/L$.

The proofs of above Lemmas are omitted here. Their proofs can be found in Rockafellar (1970); Xiao and Zhang (2014).

Proof of Main Results Using the update rule, we know

$$\begin{aligned} \|\theta_{t+1}^{s+1} - \theta_*\|^2 &= \|\|\theta_t^{s+1} - \gamma g_t^{s+1} - \theta_*\|^2\| \\ &= \|\theta_t^{s+1} - \theta_*\|^2 - 2\gamma(g_t^{s+1})^T(\theta_t - \theta_*) + \gamma^2\|g_t^{s+1}\|^2. \end{aligned} \tag{30}$$

By applying Lemma 2 with $\theta = \theta_t^{s+1}$, $v = v_t^{s+1}$, $\theta^+ = \theta_{t+1}^{s+1}$, $g = g_t^{s+1}$ and $y = \theta_*$, we get

$$-(g_t^{s+1})^T(\theta_t^{s+1} - \theta_*) + \frac{\gamma}{2}\|g_t^{s+1}\|^2 \leq P(\theta_*) - P(\theta_{t+1}^{s+1}) - \frac{\mu_F}{2}\|\theta_t^{s+1} - \theta_*\|^2 - \frac{\mu_R}{2}\|\theta_{t+1}^{s+1} - \theta_*^{s+1}\|^2 - \Delta_t^T(\theta_{t+1}^{s+1} - \theta_*),$$

where $\Delta_t^{s+1} = v_t^{s+1} - \nabla F(\theta_t^{s+1})$. Therefore,

$$\begin{aligned} \|\theta_{t+1}^{s+1} - \theta_*^{s+1}\|^2 &\leq \|\theta_t^{s+1} - \theta_*^{s+1}\|^2 - \gamma\mu_F\|\theta_t^{s+1} - \theta_*\|^2 - \gamma\mu_R\|\theta_{t+1}^{s+1} - \theta_*\|^2 \\ &\quad -2\gamma(P(\theta_{t+1}^{s+1}) - P(\theta_*)) - 2\gamma\Delta_t^T(\theta_{t+1}^{s+1} - \theta_*) \\ &\leq \|\theta_t^{s+1} - \theta_*\|^2 - 2\gamma(P(\theta_{t+1}^{s+1}) - P(\theta_*)) - 2\gamma\Delta_t^T(\theta_{t+1}^{s+1} - \theta_*). \end{aligned} \tag{31}$$

We next bound the quantity $-2\gamma(\Delta_t^{s+1})^T(\theta_{t+1}^{s+1} - \theta_*)$. We define the full proximal gradient update as

$$\bar{\theta}_{t+1}^{s+1} = prox_{\gamma R}(\theta_t^{s+1} - \gamma\nabla F(\theta_t^{s+1})),$$

though it is not used in algorithm. Then,

$$\begin{aligned} -2\gamma(\Delta_t^{s+1})^T(\theta_{t+1}^{s+1} - \theta_*) &= -2\gamma(\Delta_t^{s+1})^T(\theta_{t+1}^{s+1} - \bar{\theta}_{t+1}^{s+1}) - 2\gamma(\Delta_t^{s+1})^T(\bar{\theta}_{t+1}^{s+1} - \theta_*^{s+1}) \\ &\leq 2\gamma\|\Delta_t^{s+1}\|\|\theta_{t+1}^{s+1} - \bar{\theta}_{t+1}^{s+1}\| - 2\gamma(\Delta_t^{s+1})^T(\bar{\theta}_{t+1}^{s+1} - \theta_*) \\ &\leq 2\gamma\|\Delta_t^{s+1}\|\|(\theta_t^{s+1} - \gamma v_t^{s+1}) - (\theta_t^{s+1} - \gamma\nabla F(\theta_t^{s+1}))\| \\ &\quad -2\gamma(\Delta_t^{s+1})^T(\bar{\theta}_{t+1}^{s+1} - \theta_*) \\ &= 2\gamma^2\|\Delta_t^{s+1}\|^2 - 2\gamma(\Delta_t^{s+1})^T(\bar{\theta}_{t+1}^{s+1} - \theta_*), \end{aligned} \tag{32}$$

Thus (31) becomes

$$\|\theta_{t+1}^{s+1} - \theta_*\|^2 \leq \|\theta_t^{s+1} - \theta_*\|^2 - 2\gamma(P(\theta_{t+1}^{s+1}) - P(\theta_*)) + 2\gamma^2\|\Delta_t^{s+1}\|^2 - 2\gamma(\Delta_t^{s+1})^T(\bar{\theta}_{t+1}^{s+1} - \theta_*).$$

We take expectation on both sides with respect to i_t^{s+1} and $z_{i_t^{s+1}}$ to get

$$\mathbb{E}\|\theta_{t+1}^{s+1} - \theta_*\|^2 \leq \|\theta_t^{s+1} - \theta_*\|^2 - 2\gamma\mathbb{E}(P(\theta_{t+1}^{s+1}) - P(\theta_*)) + \gamma\eta,$$

where we use the boundness of $\mathbb{E}\|\Delta_t^{s+1}\|$ and $\|\theta_{t+1}^{s+1} - \theta_*\| \le \eta$. Before the termination of Algorithm 1, we know that $E(P(\theta_{t+1}^{s+1}) - P(\theta_*)) = \Omega(\eta)$. Therefore, we have

$$\mathbb{E}\|\theta_{t+1}^{s+1} - \theta_*\|^2 \le \|\theta_t^{s+1} - \theta_*\|^2 - 2c\gamma\mathbb{E}(P(\theta_{t+1}^{s+1}) - P(\theta_*)). \tag{33}$$

By summing the above inequality over all s and t, then we get

$$2cmT\epsilon \le \sum_{s=1}^{T}\sum_{t=1}^{m} 2c\gamma\mathbb{E}(P(\theta_{t+1}^{s+1}) - P(\theta_*)) \le \|\theta^0 - \theta_*\|^2. \tag{34}$$

We get $T(\epsilon) \le \frac{\|\theta^0 - \theta^*\|^2}{2cm\epsilon}$. This concludes the proof of Theorem 2.

Then the total computational complexity is

$$O(\frac{\|\theta_0 - \theta_*\|^2}{m\gamma\epsilon}\max\{m, n_1\})$$

for any $\epsilon = \Omega(\frac{1}{\sqrt{n_1}} + m\gamma)$. When $m = n^{\alpha_1}$, $n_1 = m^{2(\alpha-\alpha_1)}$ and $\gamma = n^{-\alpha}$ with $\alpha_1 = 2/3\alpha$, the computational complexity is $O(n^{\alpha}\|\theta_0 - \theta_*\|^2/\epsilon)$.

4. Proof of Theorem 4

Let S_1^* be the set of indices corresponding to position of θ^* where true value is non-zero and S_0^* be the set of indices corresponding to position of θ^* where true value is zero. For notational simplicity, we define $\theta_{(1)} = \theta[S_1^*]$ and $\theta_{(0)} = \theta[S_0^*]$. Next, we show that the solution $\hat{\theta}$ with $\hat{\theta}[S_0^*] = 0$ satisfies Karush-Kuhn-Tucker (KKT) condition. We then can write

$$\nabla F(\theta) = \begin{pmatrix} \nabla_1 F(\theta) \\ \nabla_0 F(\theta) \end{pmatrix},$$

and write

$$\nabla^2 F(\theta) = \begin{pmatrix} \nabla_{11}^2 F(\theta) & \nabla_{10}^2 F(\theta) \\ \nabla_{01}^2 F(\theta) & \nabla_{00}^2 F(\theta) \end{pmatrix},$$

where $\nabla_1 F(\theta)$ is the subvector of gradient corresponding to $\theta_{(1)}$ and $\nabla_{11}^2 F(\theta)$ is the block of Hessian matrix corresponding to $\theta_{(1)}$. Rest quantities are defined in the same fashion.

We then recall the irrepresentable condition.

- Assume there exists a positive constant η such that

$$|\nabla_{01} F(\theta^*)\nabla_{11}^2 F(\theta^*))^{-1}\text{sign}(\theta_{(1)}^*)| \preceq 1 - \eta. \tag{35}$$

  Here the "$\preceq$" means that the inequality holds element-wisely.

We expand $\nabla F(\hat{\theta})$ at θ^* by Taylor expansion. Then we get

$$\nabla F(\hat{\theta}) = \nabla F(\theta^*) + \nabla^2 F(\theta^*)(\hat{\theta} - \theta^*) + O((\hat{\theta} - \theta^*)^2). \tag{36}$$

For subvector $\hat{\theta}_{(1)}$ and $\theta_{(1)}^*$, we can get the similar equation, that is,

$$\nabla_1 F(\hat{\theta}) = \nabla_1 F(\theta^*) + \nabla_{11}^2 F(\theta^*)(\hat{\theta}_{(1)} - \theta_{(1)}^*) + O((\hat{\theta}_{(1)} - \theta_{(1)}^*)^2). \tag{37}$$

This implies

$$\hat{\theta}_{(1)} - \theta_{(1)}^* = -(\nabla_{11}^2 F(\theta^*))^{-1}(\nabla_1 F(\hat{\theta}) + O_p(\frac{1}{\sqrt{n}})) \tag{38}$$

For those positions in S_0^*, we can compute

$$\begin{aligned}\nabla_0 F(\hat{\theta}) &= -\nabla_{01}F(\theta^*)(\nabla_{11}^2 F(\theta^*))^{-1}(\nabla_1 F(\hat{\theta}) + O_p(\frac{1}{\sqrt{n}})) + O((\hat{\theta}-\theta^*)^2) \\ &= -\nabla_{01}F(\theta^*)(\nabla_{11}^2 F(\theta^*))^{-1}\nabla_1 F(\hat{\theta}) + O_p(\frac{1}{\sqrt{n}}) + O_p(n^{-\alpha/3}) \qquad (39)\end{aligned}$$

Note that $\nabla_1 F(\hat{\theta}) = \tau \text{sign}(\theta^*_{(1)}) + O_p(n^{-\alpha/6})$, we have

$$|\nabla_0 F(\hat{\theta})| \preceq \xi_c(\tau(1-\eta) + O_p(n^{-\alpha/6})) + O_p(\frac{1}{\sqrt{n}}) + O_p(n^{-\alpha/3}) < \tau \qquad (40)$$

when $\tau \geq n^{-\alpha/6}$. Thus, we know that $\hat{\theta}_{(0)} = \mathbf{0}$. This completes the proof.

5. Proof of Results in Network Case

Let d_i be the number of nodes that the i-th node connects to and A be the edge list. Let $|A|$ be the cardinality of A and we know $2|A| = \sum d_i$. The objective function is

$$L(\theta) = \sum_{\mathbf{z}} p(\mathbf{z}) \prod_{(i,j)\in A} f(\theta|z_i, z_j). \qquad (41)$$

Thus

$$\begin{aligned}\nabla \log L(\theta) &= \nabla \log\{\sum_{\mathbf{z}} p(\mathbf{z}) \prod_{(i,j)\in A} f_\theta(y_{ij}|z_i, z_j)\} \\ &= \sum_{\mathbf{z}} \{\nabla \log \big(p(\mathbf{z}) \prod_{(i,j)\in A} f_\theta(y_{ij}|z_i, z_j)\big)\} p(\mathbf{z}|\theta) \\ &= \sum_{\mathbf{z}} \{\nabla \log \prod_{(i,j)\in A} f_\theta(y_{ij}|z_i, z_j)\} p(\mathbf{z}|\theta) \qquad (42) \\ &= \sum_{\mathbf{z}} \{\sum_{(i,j)\in A} \nabla \log f_\theta(y_{ij}|z_i, z_j)\} p(\mathbf{z}|\theta) \qquad (43) \\ &= \mathbb{E}_{\mathbf{z}}\{\sum_{(i,j)\in A} \nabla \log f_\theta(y_{ij}|z_i, z_j)\}. \qquad (44)\end{aligned}$$

Next we show the local convergence property of Algorithm 2 under the latent network setting. For any $\theta \in B(\theta^*, \delta)$ with some small radius δ, we can show that $p(\mathbf{z}|\theta) \to \mathbf{1}_{\mathbf{z}=\mathbf{z}^*}$ (i.e., a probability mass function which puts total point probability on true latent memberships). More specifically, according to Lemma 3, it gives $d_{TV}(p(\mathbf{z}|\theta), \mathbf{1}_{\mathbf{z}=\mathbf{z}^*}) \leq \exp\{-cd_{min}\}$ for some positive constant c and d_{min} is the minimum of d_i's.

We first prove several useful lemmas.

Lemma 3 *For any $\theta \in B(\theta^*, \delta)$, there exists a constant c such that*

$$\|p(\mathbf{z}|\theta) - \mathbf{1}_{\mathbf{z}=\mathbf{z}^*}\|_{TV} \leq \exp\{-cd_{min}\}. \qquad (45)$$

Proof of Lemma 3 To prove (3), it is equivalent to prove

$$\sum_{\mathbf{z}\neq\mathbf{z}^*} p_\theta(\mathbf{z}) = p_\theta(\mathbf{z}^*) \cdot \exp\{-cd_{min}\}, \qquad (46)$$

where $p_\theta(\mathbf{z}) = p_\theta(\mathbf{z}, \mathbf{y})$ is the complete likelihood function. We omit script $\mathbf{y}$ for notational simplicity.

The main step of the proof is to show that

$$\log p_\theta(\mathbf{z}) \leq \log p_\theta(\mathbf{z}^*) - cd_{min}|\mathbf{z} - \mathbf{z}^*|_0 \qquad (47)$$

holds for all $\mathbf{z} \neq \mathbf{z}^*$ with high probability. According to concentration lemma 5, we have

$$\begin{aligned} & P(|\log p_\theta(\mathbf{z}) - \log p_\theta(\mathbf{z}^*) - \mathbb{E}[\log p_\theta(\mathbf{z}) - \log p_\theta(\mathbf{z}^*)]| \geq |\mathbf{z} - \mathbf{z}^*|_0 d_{min} x) \\ \leq \quad & \exp\{-|\mathbf{z} - \mathbf{z}^*|_0 d_{min} x^2\} \end{aligned} \tag{48}$$

by taking $g_\theta(z) = \log p_\theta(\mathbf{z}) - \log p_\theta(\mathbf{z}^*)$. By model identifiability, we know that there exists constant c_0 such that

$$\mathbb{E}[\log p_\theta(\mathbf{z})] - \mathbb{E}[\log p_\theta(\mathbf{z}^*)] \leq c_0 |\mathbf{z} - \mathbf{z}^*|_0 d_{min}. \tag{49}$$

By taking $x = c_0/2$ in (48), we have

$$\log p_\theta(\mathbf{z}) \leq \log p_\theta(\mathbf{z}^*) - \frac{c_0}{2} d_{min} |\mathbf{z} - \mathbf{z}^*|_0 \tag{50}$$

with probability at least $1 - \exp\{-|\mathbf{z} - \mathbf{z}^*|_0 d_{min} c_0^2/4\}$. Therefore, we have

$$\begin{aligned} & P(\log p_\theta(\mathbf{z}) \leq \log p_\theta(\mathbf{z}^*) - \frac{c_0}{2} d_{min} |\mathbf{z} - \mathbf{z}^*|_0, \text{ for any } \mathbf{z}) \\ \geq \quad & 1 - \sum_{\mathbf{z}} \exp\{-|\mathbf{z} - \mathbf{z}^*|_0 dv c_0^2/4\} \\ = \quad & 1 - \sum_{d=1}^{n} \sum_{\mathbf{z}:|\mathbf{z}-\mathbf{z}^*|_0} \exp\{-|\mathbf{z} - d dv c_0^2/4\} \\ = \quad & 1 - \sum_{d=1}^{n} \frac{n!}{(n-d)!(d)!} \exp\{-d dv c_0^2/4\} \\ \geq \quad & 1 - \sum_{d=1}^{n} n^d \exp\{-d d_{min} c_0^2/4\} && (51) \\ \geq \quad & 1 - (1 - \exp\{-d_{min} c_0^2/4 + \log n\})^{-1} \exp\{-d_{min} c_0^2/4 + \log n\} \\ \geq \quad & 1 - \exp\{-c' d_{min}\} && (52) \end{aligned}$$

by adjusting the constant c' and the fact that $d_{min} \gg \log n$. This establishes (46) and the lemma follows as well.

Lemma 4 *For any $\theta \in B(\theta^*, \delta)$, there exist constants c', c'' such that*

$$\|\nabla L(\theta) - \nabla L(\theta|\mathbf{z}^*)\| \leq \exp\{-c' d_{min}\}. \tag{53}$$

and

$$\|\nabla L_i(\theta) - \nabla L_i(\theta|\mathbf{z}^*)\| \leq \exp\{-c' d_{min}\} \tag{54}$$

hold with probability at least $1 - \exp\{-c'' d_{min}\}$.

The proof of Lemma 4 is similar to that of Lemma 3. Hence, we omit here.

Lemma 5 *Suppose $g_\theta(z)$ is any function of form $\sum_{i \in A_s} \sum_{j \in A_i} \log f_\theta(y_{ij}|z_i, z_j)$, where A_s is arbitrary subset of $\{1, \ldots, n\}$. Then it holds that*

$$P(|g_\theta(z) - \mathbb{E} g_\theta(z)| \geq |\bar{A}_s| x) \leq \exp\{-C|A_s| d_{min} x^2\}, \tag{55}$$

for some constant C. Here $\bar{A}_s := \{(i,j) : i \in A_s, j \in A_i\}$.

Proof of Lemma 5 By boundness assumption, we know there exist constant M such that $|\log f_\theta(y_{ij}|z_i, z_j) - \mathbb{E} \log f_\theta(y_{ij}|z_i, z_j)| \leq M$. Then, by Hoeffding's inequality, we have

$$\begin{aligned} P(|g_\theta(z) - \mathbb{E} g_\theta(z)| \geq |\bar{A}_s| x) \quad & \leq \quad \exp\{-2 \frac{|\bar{A}_s|^2 x^2}{|\bar{A}_s| M^2}\}, \\ & \leq \quad \exp\{-2 \frac{|\bar{A}_s| x^2}{M^2}\} \\ & \leq \quad \exp\{-C|A_s| d_{min} x^2\} && (56) \end{aligned}$$

by adjusting the constant C. This concludes the lemma.

We define the following quantities,

$$\nabla H(\theta, \mathbf{z}) := \frac{1}{|A|} \sum_{(i,j)\in A} \nabla \log f_\theta(y_{ij}|z_i, z_j),$$

$$\nabla H_i(\theta, \mathbf{z}) := \frac{1}{d_i} \sum_{j:(i,j)\in A} \nabla \log f_\theta(y_{ij}|z_i, z_j),$$

and

$$\nabla H_B(\theta, \mathbf{z}) := \frac{1}{\sum_{i\in B} d_i} \sum_{(i,j)\in A, i\in B} \nabla \log f_\theta(y_{ij}|z_i, z_j).$$

We also define

$$\nabla H(\theta) := \frac{1}{|A|} \sum_{(i,j)\in A} \nabla \log f_\theta(y_{ij}|z_i^*, z_j^*),$$

$$\nabla H_i(\theta) := \frac{1}{d_i} \sum_{j:(i,j)\in A} \nabla \log f_\theta(y_{ij}|z_i^*, z_j^*),$$

and

$$\nabla H_B(\theta) := \frac{1}{\sum_{i\in B} d_i} \sum_{(i,j)\in A, i\in B} \nabla \log f_\theta(y_{ij}|z_i^*, z_j^*).$$

Therefore, we can compute

$$\begin{aligned}
\mathbb{E}_{i_t^{s+1}, z_{i_t^{s+1}}} v_t^{s+1} &= \frac{1}{2|A|} \sum_{i=1}^n \mathbb{E}_{z_i \sim p_{\theta_t^{s+1}}(z|\mathbf{y})} \nabla H_i(\theta_t^{s+1}, \mathbf{z}_t^{s+1}) \\
&\quad - \frac{1}{2|A|} \sum_{i=1}^n \mathbb{E}_{z_i \sim p_{\theta_t^{s+1}}(z|\mathbf{y})} \nabla H_i(\tilde\theta^s, \mathbf{z}_t^{s+1}) + \nabla H_B(\tilde\theta^s, \mathbf{z}^s) \\
&= \nabla H(\theta_t^{s+1}) - \nabla H(\tilde\theta^s) + \nabla H_B(\tilde\theta^s) + O(\exp\{-c' d_{min}\}), \\
&:= H_t^{s+1}
\end{aligned} \tag{57}$$

according to Lemma 3. We next consider to compute the upper bound of $\mathbb{E}[\|v_t^{s+1}\|^2]$

$$\begin{aligned}
\mathbb{E}[\|v_t^{s+1}\|^2] &= \mathbb{E}[\|v_t^{s+1} - H_t^{s+1} + H_t^{s+1}\|^2] \\
&\le 2\mathbb{E}[\|H_t^{s+1}\|^2] + 2\mathbb{E}[\|\xi_t^{s+1} - \mathbb{E}[\xi_t^{s+1}]\|^2] \\
&\le 2\mathbb{E}[\|H_t^{s+1}\|^2] + 2\mathbb{E}[\|\xi_t^{s+1}\|^2], \\
&\le 4\mathbb{E}[\|\nabla \log L(\theta_t^{s+1})\|^2] + 4\eta^2 + 2\mathbb{E}[\|\xi_t^{s+1}\|^2], \\
&\le 4C\mathbb{E}[\|\nabla \log L(\theta_t^{s+1})\|^2] + 2\mathbb{E}[\|\xi_t^{s+1}\|^2],
\end{aligned} \tag{58}$$

where $\xi_t^{s+1} = \frac{1}{d_i}\{\nabla H_{i_t}(\theta_t^{s+1}, z_{i_t}) - \nabla H_{i_t}(\tilde\theta^s, z_{i_t})\}$ and $\eta = \nabla H_t^{s+1} - \nabla \log L(\theta_t^{s+1})$ is the step error which is order of $(m\gamma + \frac{1}{\sqrt{n_1 n}})$. The last inequality above uses the fact that $\|\nabla \log L(\theta_t^{s+1})\| = \Omega(\|\eta\|)$ before the termination of the algorithm. Since, $\|\eta\|$ can be bounded by

$$\begin{aligned} \|\eta\| &= \|\nabla H_t^{s+1} - \nabla \log L(\theta_t^{s+1})\| \\ &= \|\nabla H(\theta_t^{s+1}) - \nabla H(\tilde{\theta}^s) + \nabla H_{B^s}(\tilde{\theta}^s) - \nabla \log L(\theta_t^{s+1})\| + O(\exp\{-c' d_{min}\}) \\ &\leq \|\nabla H(\theta_t^{s+1}) - \nabla H(\tilde{\theta}^s)\| + \|\nabla H_{B^s}(\tilde{\theta}^s) - \nabla \log L(\tilde{\theta}^s|\mathbf{z}^*)\| \\ &\quad + \|\nabla \log L(\tilde{\theta}^s|\mathbf{z}^*) - \nabla \log L(\theta_t^{s+1}|\mathbf{z}^*)\| \\ &\quad + \|\nabla \log L(\theta_t^{s+1}|\mathbf{z}^*) - \nabla \log L(\theta_t^{s+1})\| + O(\exp\{-c' d_{min}\}) \\ &\leq L\|\theta_t^{s+1} - \tilde{\theta}^s\| + O(\frac{1}{\sqrt{d_{min}|B^s|}}) \\ &\quad + L\|\theta_t^{s+1} - \tilde{\theta}^s\| + O(\exp\{-c' d_{min}\}) + O(\exp\{-c' d_{min}\}) \qquad (59) \\ &= O(m\gamma + \frac{1}{\sqrt{n_1 d_{min}}}). \qquad (60) \end{aligned}$$

Here, (59) uses the fact that $\nabla H(\theta)$ and $\nabla \log L(\theta|\mathbf{z}^*)$ are L-Lipschitz continuous for some L and $\|\nabla \log H_{B^s}(\tilde{\theta}^s|\mathbf{z}^*) - \nabla \log L(\tilde{\theta}^s|\mathbf{z}^*)\|$ is $O_p(1/\sqrt{d_{min}|B^s|})$ by using concentration inequality 5.

As a result, we can obtain

$$\begin{aligned} \langle \nabla \log L(\theta_t^{s+1}), H_t^{s+1} \rangle &\geq \langle \nabla \log L(\theta_t^{s+1}), \nabla \log L(\theta_t^{s+1}) \rangle - \langle \nabla \log L(\theta_t^{s+1}), \nabla \log L(\theta_t^{s+1}) - H_t^{s+1} \rangle \\ &\geq \|\nabla \log L(\theta_t^{s+1})\|^2 - \|\nabla \log L(\theta_t^{s+1})\| \|\nabla \log L(\theta_t^{s+1}) - H_t^{s+1}\| \\ &\geq c\|\nabla \log L(\theta_t^{s+1})\|^2, \qquad (61) \end{aligned}$$

when $\|\nabla \log L(\theta_t^{s+1})\| = \Omega(m\gamma + \frac{1}{\sqrt{n_1 d_{min}}})$ before the termination of the algorithm.

Under the network setting, we can similarly construct the Lyapunov function as

$$R_t^{s+1} := \mathbb{E}[F(\theta_t^{s+1}) + c_t\|\theta_t^{s+1} - \tilde{\theta}^s\|^2],$$

with c_t's satisfying recursive relationship $c_t = c_{t+1}(1 + \gamma\beta + 2\gamma^2 L^2) + \gamma^2 L^3$ $(t = 0, \ldots, m-1)$ and $c_m = 0$. By the same procedure, we then arrive at

$$\frac{1}{T}\sum_{s=0}^{T-1}\sum_{t=0}^{m-1} \mathbb{E}[\|\nabla F(\theta_t^{s+1})\|^2] \leq C\frac{R_0^0 - R_m^T}{\gamma m T} \leq C\frac{F(\theta^0) - F(\theta^*)}{\gamma m T} \qquad (62)$$

holds for some constant C. This leads to the desire result and concludes the proof of Theorem 5.

Finally, we set $n_1 = n^{2(\alpha-\alpha_1)}/d_{min}$, $m = n^{\alpha_1}$, $\gamma = n^{-\alpha}$. Then the total computational complexity will be

$$(md_{max} + n_1 d_{max})\frac{C}{\gamma m \epsilon}. \qquad (63)$$

Suppose $d_{min}, d_{max} \approx n^{\alpha_0}$, then we can choose $\alpha_1 = \frac{2\alpha-\alpha_0}{3}$. Then $n_1 = n^{(2\alpha-\alpha_0)/3}$ and computational complexity becomes $n^{\alpha+\alpha_0}\frac{C}{\epsilon}$, where α should satisfy $\alpha < 1$ and $2(\alpha - \alpha_1) > \alpha_0$ (i.e., $\alpha > \alpha_0/2$).

References

Sashank J. Reddi, Ahmed Hefny, Suvrit Sra, Barnabás Póczos, and Alexander J. Smola. Stochastic variance reduction for nonconvex optimization. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pages 314–323, New York City, NY, 2016.

R Tyrrell Rockafellar. *Convex Analysis*, volume 36. Princeton university press, 1970.

Lin Xiao and Tong Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.