Supplementary Materials

for "Estimation of Markov Chain via Rank-constrained Likelihood"

Xudong Li [1] **Mengdi Wang** [1] **Anru Zhang** [2]

1. Proof of Proposition 1

Proof. Given $x_k = i$, x_{k+1} is with discrete distribution $\mathbf{P}_{i\cdot}$. Thus, the log-likelihood of $x_{k+1}|x_k = \log(\mathbf{P}_{x_k,x_{k+1}}) = \langle \mathbf{P}, e_{x_k} e_{x_{k+1}}^\top \rangle$. Then the negative log-likelihood given $\{x_0, \ldots, x_n\}$ is

$$-\sum_{k=1}^{n} \log(\mathbf{P}_{x_k,x_{k+1}}) = \langle \log(\mathbf{P}), e_{x_k} e_{x_{k+1}}^\top \rangle = -\sum_{i=1}^{p}\sum_{j=1}^{p} n_{ij} \log(\mathbf{P}_{ij}).$$

□

2. Proof of Theorem 1

Proof. Recall $D_{KL}(\mathbf{P}, \mathbf{Q}) = \sum_{i=1}^{p} \mu_i D_{KL}(P_{i\cdot}, Q_{i\cdot}) = \sum_{j=1}^{p} \mu_i P_{ij} \log(P_{ij}/Q_{ij})$. For convenience, we also denote,

$$\tilde{D}(\mathbf{P}, \mathbf{Q}) = \frac{1}{n}\sum_{k=1}^{n} \langle \log(\mathbf{P}) - \log(\mathbf{Q}), \mathbf{E}_k \rangle,$$

where $\mathbf{E}_k = e_i e_j^\top$ if the k-th jump is from States i to j. Then $(\mathbf{E}_k)_{k=1}^n$ be independent copies such that $P(\mathbf{E}_k = e_i e_j^\top) = \mu_i P_{ij}$, and

$$L(\mathbf{P}) = -\frac{1}{n}\sum_{i,j=1}^{p} n_{ij}\log(P_{ij}) = -\frac{1}{n}\sum_{k=1}^{n} \log\langle \mathbf{X}, \mathbf{E}_k \rangle$$

By the property of the programming,

$$\tilde{D}(\mathbf{P}, \hat{\mathbf{P}}) = \frac{1}{n}\sum_{k=1}^{n} \langle \log(\mathbf{P}) - \log(\hat{\mathbf{P}}), \mathbf{E}_k \rangle = L(\hat{\mathbf{P}}) - L(\mathbf{P}) \leq 0. \tag{1}$$

Based on the assumption, $\mathrm{rank}(\mathbf{P}) \wedge \mathrm{rank}(\hat{\mathbf{P}}) \leq r$. For any $\mathbf{Q}$ with $\mathrm{rank}(\mathbf{Q}) \leq r$, we must have $\mathrm{rank}(\mathbf{Q} - \mathbf{P}) \leq 2r$. Due to the duality between operator and spectral norm,

$$\|\mathbf{Q} - \mathbf{P}\|_* \leq \sqrt{2r}\|\mathbf{Q} - \mathbf{P}\|_F. \tag{2}$$

Next, we denote $\eta = C_\eta\sqrt{\log p/n}$ for some large constant $C_\eta > 0$, and introduce the following deterministic set in $\mathbb{R}^{p\times p}$,

$$\mathcal{C} = \left\{\mathbf{Q} : \alpha/p \leq Q_{ij} \leq \beta/p,\ \mathrm{rank}(Q) \leq r, D_{KL}(\mathbf{P}, \mathbf{Q}) \geq \eta \right\}.$$

[1]Department of Operations Research and Financial Engineering, Princeton University, Sherrerd Hall, Princeton, NJ 08544 [2]Department of Statistics, University of Wisconsin-Madison, Madison, WI 53706. Correspondence to: Xudong Li <xudongl@princeton.edu>, Mengdi Wang <mengdiw@princeton.edu>, Anru Zhang <anruzhang@stat.wisc.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

We particularly aim to show next that

$$P\left\{\forall \mathbf{Q}\in\mathcal{C},\quad \left|\tilde{D}(\mathbf{P},\mathbf{Q})-D_{KL}(\mathbf{P},\mathbf{Q})\right|\leq \frac{1}{2}D_{KL}(\mathbf{P},\mathbf{Q})+\frac{Cpr\log(p)}{n}\right\}\geq 1-Cp^{-c}. \tag{3}$$

In order to prove (3), we first split $\mathcal{C}$ as the union of the sets,

$$\mathcal{C}_l=\left\{\mathbf{Q}\in\mathcal{C}: 2^{l-1}\eta\leq D_{KL}(\mathbf{P},\mathbf{Q})\leq 2^l\eta,\ \mathrm{rank}(Q)\leq r\right\},\quad l=1,2,3,\ldots.$$

where η is to be determined later. Define

$$\begin{aligned}\gamma_l&=\sup_{\mathbf{Q}\in\mathcal{C}_l}\left|D_{KL}(\mathbf{P},\mathbf{Q})-\tilde{D}(\mathbf{P},\mathbf{Q})\right|\\ &=\sup_{\mathbf{Q}\in\mathcal{C}_l}\left|\frac{1}{n}\sum_{k=1}^n\langle\log(\mathbf{P})-\log(\mathbf{Q}),\mathbf{E}_k\rangle-\mathbb{E}\langle\log(\mathbf{P})-\log(\mathbf{Q}),\mathbf{E}_k\rangle\right|.\end{aligned}$$

Since $|\log(P_{ij})-\log(Q_{ij})|\leq\log(\beta/\alpha)$, we apply a empirical process version of Hoeffding's inequality (Theorem 14.2 in (Bühlmann & Van De Geer, 2011)),

$$P\left(\gamma_l-\mathbb{E}(\gamma_l)\geq 2^{l-3}\cdot\eta\right)\leq\exp\left(-\frac{cn\cdot 4^{l-3}\eta^2}{(\log(\beta/\alpha))^2}\right). \tag{4}$$

for constant $c>0$. We generate $\{\varepsilon_k\}_{k=1}^n$ as i.i.d. Rademacher random variables. By a symmetrization argument in empirical process,

$$\begin{aligned}\mathbb{E}\gamma_l=&\mathbb{E}\left(\sup_{\mathbf{Q}\in\mathcal{C}_l}\left|\frac{1}{n}\sum_{k=1}^n\langle\log\mathbf{P}-\log\mathbf{Q},\mathbf{E}_k\rangle-\mathbb{E}\frac{1}{n}\sum_{k=1}^n\langle\log\mathbf{P}-\log\mathbf{Q},\mathbf{E}_k\rangle\right|\right)\\ \leq&2\mathbb{E}\left(\sup_{\mathbf{Q}\in\mathcal{C}_l}\left|\frac{1}{n}\sum_{k=1}^n\varepsilon_k\langle\log\mathbf{P}-\log\mathbf{Q},\mathbf{E}_k\rangle\right|\right).\end{aligned}$$

Let $\phi_k(t)=\alpha/p\cdot\langle\log(\mathbf{P})-\log(\mathbf{Q}+t),\mathbf{E}_k\rangle$, then $\phi_k(0)=0$ and $|\phi_k'(t)|\leq 1$ for all t if $t+P_{ij}\geq\alpha/p$. In other words, $\phi_{k,i,j}$ is a contraction map for $t\geq\min_{i,j}(P_{ij}-\alpha/p)$. By concentration principle (Theorem 4.12 in (Ledoux & Talagrand, 2013)),

$$\begin{aligned}\mathbb{E}(\gamma_l)\leq&\frac{2p}{\alpha}\mathbb{E}\left(\sup_{\mathbf{Q}\in\mathcal{C}_l}\left|\frac{1}{n}\sum_{k=1}^n\varepsilon_k\phi_k\left(\langle\mathbf{Q}-\mathbf{P},\mathbf{E}_k\rangle\right)\right|\right)\\ \leq&\frac{4p}{\alpha}\mathbb{E}\left(\sup_{\mathbf{Q}\in\mathcal{C}_l}\left|\frac{1}{n}\sum_{k=1}^n\varepsilon_k\langle\mathbf{Q}-\mathbf{P},\mathbf{E}_k\rangle\right|\right)\\ \leq&\frac{4p}{\alpha}\mathbb{E}\left(\sup_{\mathbf{Q}\in\mathcal{C}_l}\left\|\frac{1}{n}\sum_{k=1}^n\varepsilon_k\mathbf{E}_k\right\|\cdot\|\mathbf{Q}-\mathbf{P}\|_*\right)\\ \leq&\frac{4p}{\alpha}\mathbb{E}\left\|\frac{1}{n}\sum_{k=1}^n\varepsilon_k\mathbf{E}_k\right\|\cdot\sup_{\mathbf{Q}\in\mathcal{C}_l}\|\mathbf{Q}-\mathbf{P}\|_*\end{aligned} \tag{5}$$

By $\mathrm{rank}(\mathbf{P})\wedge\mathrm{rank}(\mathbf{Q})\leq r$ and Lemma 5 in (Zhang & Wang, 2018),

$$\begin{aligned}&\sup_{\mathbf{Q}\in\mathcal{C}_l}\|\mathbf{Q}-\mathbf{P}\|_*\overset{(2)}{\leq}\sup_{\mathbf{Q}\in\mathcal{C}_l}\sqrt{2r}\|\mathbf{Q}-\mathbf{P}\|_F\\ \leq&\sqrt{\frac{r(\beta/p)^2}{(\alpha/p)}\sum_{i=1}^p D(P_{i\cdot},Q_{i\cdot})}\leq\sqrt{\frac{r\beta^2}{\alpha^2}\cdot 2^l\eta}.\end{aligned} \tag{6}$$

Then we evaluate $\mathbb{E}\|\frac{1}{n}\sum_{k=1}^n \varepsilon_k \mathbf{E}_k\|$. Note that $\|\mathbf{E}_k\| \leq 1$,

$$\begin{aligned}
\|\sum_{k=1}^n \mathbb{E}\mathbf{E}_k^\top \mathbf{E}_k\| =& n\left\|\sum_{i=1}^p\sum_{j=1}^p \mu_i P_{ij}(e_ie_j^\top)^\top(e_ie_j^\top)\right\| = n\left\|\sum_{j=1}^p (\mu^\top P)_j e_je_j^\top\right\| \\
=& n\left\|\sum_{j=1}^p \mu_j e_je_j^\top\right\| \leq n\mu_{\max};
\end{aligned}$$

$$\begin{aligned}
\|\sum_{k=1}^n \mathbb{E}\mathbf{E}_k \mathbf{E}_k^\top\| =& n\left\|\sum_{i=1}^p\sum_{j=1}^p \mu_i P_{ij}(e_ie_j^\top)(e_ie_j^\top)^\top\right\| = \left\|\sum_{i=1}^p\sum_{j=1}^p \mu_i P_{ij} e_ie_i^\top\right\| \\
=& \left\|\sum_{j=1}^p \mu_j e_je_j^\top\right\| \leq n\mu_{\max}.
\end{aligned}$$

By Theorem 1 in (Tropp, 2016),

$$\mathbb{E}\left\|\frac{1}{n}\sum_{k=1}^n \varepsilon_k \mathbf{E}_k\right\| \leq \frac{C\sqrt{n\mu_{\max}\log p}}{n} + \frac{C\log p}{n} \leq C\sqrt{\frac{\mu_{\max}\log p}{n}} \leq \sqrt{\frac{\beta\log p}{np}}. \tag{7}$$

provided that $n \geq Cp\log(p)$. Combining (4), (5), (6), and (7), we have

$$\mathbb{E}\gamma_l \leq C\sqrt{\frac{pr\log p}{n}\cdot 2^l\eta} \leq C^2\frac{pr\log p}{2n} + 2^{l-3}\eta,$$

$$P\left(\gamma_l \geq 2^{l-2}\eta + \frac{Cpr\log p}{n}\right) \leq \exp\left(-cn\cdot 4^l\eta^2\right).$$

Now,

$$\begin{aligned}
& P\left(\exists \mathbf{Q}\in\mathcal{C},\quad \left|\tilde{D}(\mathbf{P},\mathbf{Q}) - D_{KL}(\mathbf{P},\mathbf{Q})\right| > \frac{1}{2}D_{KL}(\mathbf{P},\mathbf{Q}) + \frac{Cpr\log(p)}{n}\right) \\
\leq & \sum_{l=0}^\infty P\left(\exists \mathbf{Q}\in\mathcal{C}_l,\quad \left|\tilde{D}(\mathbf{P},\mathbf{Q}) - D_{KL}(\mathbf{P},\mathbf{Q})\right| > \frac{1}{2}D_{KL}(\mathbf{P},\mathbf{Q}) + \frac{Cpr\log(p)}{n}\right) \\
\leq & \sum_{l=0}^\infty P\left(\exists \mathbf{Q}\in\mathcal{C}_l,\quad \gamma_l > 2^{l-2}\eta + \frac{Cpr\log(p)}{n}\right) \\
\leq & \sum_{l=0}^\infty \exp(-c\cdot C_\eta\cdot 4^l\log p) \leq \exp(-c\cdot C_\eta l\log(p)) \leq Cp^{-c}
\end{aligned}$$

provided reasonably large $C_\eta > 0$. Thus, we have obtained (3).

Finally, it remains to bound the errors for $\|\hat{\mathbf{P}} - \mathbf{P}\|_F$ and $D_{KL}(\mathbf{P},\hat{\mathbf{P}})$ given (3). In fact, provided that (3) holds,

- if $\hat{\mathbf{P}} \notin \mathcal{C}$, we have $D_{KL}(\mathbf{P},\hat{\mathbf{P}}) \leq C\sqrt{\frac{\log p}{n}}$;
- if $\hat{\mathbf{P}} \in \mathcal{C}$, by (3),

$$D_{KL}(\mathbf{P},\hat{\mathbf{P}}) \leq \tilde{D}(\mathbf{P},\hat{\mathbf{P}}) + \frac{Cpr\log p}{n} \overset{(1)}{\leq} \frac{Cpr\log p}{n}.$$

To sum up, we must have

$$D_{KL}(\mathbf{P},\hat{\mathbf{P}}) \leq C\sqrt{\frac{\log p}{n}} + \frac{Cpr\log p}{n}.$$

with probability at least $1 - Cp^{-c}$. For Frobenius norm error, we shall note that

$$\begin{aligned}\|\hat{\mathbf{P}} - \mathbf{P}\|_F^2 \leq \sum_{i=1}^p \|P_{i\cdot} - \hat{P}_{i\cdot}\|_2^2 \leq \sum_{i=1}^p \frac{2\beta^2}{\alpha p} D_{KL}(P_{i\cdot}, \hat{P}_{i\cdot}) \\ \leq \sum_{i=1}^p \frac{2\beta^2}{\alpha^2} \mu_i D_{KL}(P_{i\cdot}, \hat{P}_{i\cdot}) = \frac{\beta^2}{\alpha^2} D_{KL}(\mathbf{P}, \hat{\mathbf{P}}).\end{aligned}$$

Therefore, we have finished the proof for Theorem 1. □

3. Proof of Theorem 2

Proof. Based on the proof of Theorem 1 in (Zhang & Wang, 2018), one has

$$\inf_{\widehat{\mathbf{P}}} \sup_{\mathbf{P} \in \bar{\mathcal{P}}} \frac{1}{p} \sum_{i=1}^p \mathbb{E}\|\widehat{P}_{i\cdot} - P_{i\cdot}\|_1 \geq c\left(\sqrt{\frac{rp}{n}} \wedge 1\right),$$

where $\bar{\mathcal{P}} = \{\mathbf{P} \in \mathcal{P} : 1/(2p) \leq P_{ij} \leq 3/(2p)\} \subseteq \mathcal{P}$. By Cauchy Schwarz inequality,

$$\sum_{i=1}^p \|\widehat{P}_{i\cdot} - P_{i\cdot}\|_1 = \sum_{i,j=1}^p |\widehat{P}_{ij} - P_{ij}| \leq p \sqrt{\sum_{i,j=1}^p (\widehat{P}_{ij} - P_{ij})^2},$$

Thus,

$$\inf_{\widehat{\mathbf{P}}} \sup_{\mathbf{P} \in \mathcal{P}} \mathbb{E} \sum_{i=1}^p \|\widehat{P}_{i\cdot} - P_{i\cdot}\|_2^2 \geq \left(\inf_{\widehat{\mathbf{P}}} \sup_{\mathbf{P} \in \bar{\mathcal{P}}} \mathbb{E} \sum_{i=1}^p \frac{1}{p} \|\widehat{P}_{i\cdot} - P_{i\cdot}\|_1\right)^2 \geq c\left(\frac{rp}{n} \wedge 1\right) \geq \frac{cpr}{n}.$$

The lower bound for KL divergence essentially follows due to the inequalities between ℓ_2 and KL-divergence for bounded vectors in Lemma 5 of (Zhang & Wang, 2018). □

4. Proof of Theorem 3

Proof. Let $\hat{\mathbf{U}}_\perp, \hat{\mathbf{V}}_\perp \in \Re^{p \times (p-r)}$ be the orthogonal complement of $\hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$. Since $\mathbf{U}, \mathbf{V}, \hat{\mathbf{U}}$, and $\hat{\mathbf{V}}$ are the leading left and right singular vectors of $\mathbf{P}$ and $\hat{\mathbf{P}}$, we have

$$\|\hat{\mathbf{P}} - \mathbf{P}\|_F \geq \|\hat{\mathbf{U}}_\perp^\top(\hat{\mathbf{P}} - \mathbf{U}\mathbf{U}^\top\mathbf{P})\|_F = \|\hat{\mathbf{U}}_\perp^\top\mathbf{U}\mathbf{U}^\top\mathbf{P}\|_F \geq \|\hat{\mathbf{U}}_\perp^\top\mathbf{U}\|_F \cdot \sigma_r(\mathbf{U}^\top\mathbf{P}) = \|\sin\Theta(\hat{\mathbf{U}}, \mathbf{U})\|_F \cdot \sigma_r(\mathbf{P}).$$

Similar argument also applies to $\|\sin\Theta(\hat{\mathbf{V}}, \mathbf{V})\|$. Thus,

$$\max\{\|\sin\Theta(\hat{\mathbf{U}}, \mathbf{U})\|_F, \|\sin\Theta(\hat{\mathbf{V}}, \mathbf{V})\|_F\} \leq \min\left\{\frac{\|\hat{\mathbf{P}} - \mathbf{P}\|_F}{\sigma_r(\mathbf{P})}, \sqrt{r}\right\}.$$

The rest of the proof immediately follows from Theorem 1. □

5. Proof of Proposition 2

Proof. Since $\operatorname{rank}(\mathbf{X}_c^*) \leq r$, we know that $\mathbf{X}_c^*$ is in fact a feasible solution to the original problem (5) and $\|\mathbf{X}_c^*\|_* - \|\mathbf{X}_c^*\|_{(r)} = 0$. Therefore, for any feasible solution X to (5), it holds that

$$\begin{aligned}f(\mathbf{X}_c^*) = f(\mathbf{X}_c^*) + c(\|\mathbf{X}_c^*\|_* - \|\mathbf{X}_c^*\|_{(r)}) \\ \leq f(\mathbf{X}) + c(\|\mathbf{X}\|_* - \|\mathbf{X}\|_{(r)}) = f(\mathbf{X}).\end{aligned}$$

This completes the proof of the proposition. □

6. Proof of Theorem 5 (Convergence of sGS-ADMM)

Proof. In order to use (Li et al., 2016b, Theorem 3), we need to write problem (**D**) as following

$$\begin{array}{ll} \min & f^*(-\boldsymbol{\Xi}) - \langle b,\, y\rangle + \delta(\mathbf{S} \mid \|\mathbf{S}\|_2 \le c) + \frac{\alpha}{2}\|\mathbf{Z}\|_F^2 \\ \text{s.t.} & \mathcal{F}(\boldsymbol{\Xi}) + \mathcal{A}_1^*(y) + \mathcal{G}(\mathbf{S}) + \mathcal{B}_1^*(\mathbf{Z}) = \mathbf{W}, \end{array}$$

where $\mathcal{F}, \mathcal{A}_1, \mathcal{G}$ and $\mathcal{B}_1$ are linear operators such that for all $(\boldsymbol{\Xi}, y, \mathbf{S}, \mathbf{Z}) \in \Re^{p\times p} \times \Re^n \times \Re^{p\times p} \times \Re^{p\times p}$, $\mathcal{F}(\boldsymbol{\Xi}) = \boldsymbol{\Xi}$, $\mathcal{A}_1^*(y) = \mathcal{A}^*(y)$, $\mathcal{G}(\mathbf{S}) = \mathbf{S}$ and $\mathcal{B}_1^*(\mathbf{Z}) = \alpha\mathbf{Z}$. Clearly, $\mathcal{F} = \mathcal{G} = \mathcal{I}$ and $\mathcal{B}_1 = \alpha\mathcal{I}$ where $\mathcal{I} : \Re^{p\times p} \to \Re^{p\times p}$ is the identity map. Therefore, we have $\mathcal{A}_1\mathcal{A}_1^* \succ 0$ and $\mathcal{F}\mathcal{F}^* = \mathcal{G}\mathcal{G}^* = \mathcal{I} \succ 0$. Note that if $\alpha > 0$, $\mathcal{B}_1\mathcal{B}_1^* = \alpha^2\mathcal{I} \succ 0$. Hence, the assumptions and conditions in (Li et al., 2016b, Theorem 3) are satisfied whenever $\alpha \ge 0$. The convergence results thus follow directly. □

7. Proof of Theorems 4 and 6

We only need to prove Theorem 6 as Theorem 4 is a special incidence. To prove Theorem 6,we first introduce the following lemma.

Lemma 1. *Suppose that $\{x^k\}$ is the sequence generated by Algorithm 3. Then $\theta(x^{k+1}) \le \theta(x^k) - \frac{1}{2}\|x^{k+1} - x^k\|^2_{\mathcal{G}+2\mathcal{T}}$.*

Proof. For any $k \ge 0$, by the optimality condition of problem (10) at x^{k+1}, we know that there exist $\eta^{k+1} \in \partial p(x^{k+1})$ such that

$$0 = \nabla g(x^k) + (\mathcal{G} + \mathcal{T})(x^{k+1} - x^k) + \eta^{k+1} - \xi^k = 0.$$

Then for any $k \ge 0$, we deduce

$$\begin{array}{rl} & \theta(x^{k+1}) - \theta(x^k) \le \widehat{\theta}(x^{k+1}; x^k) - \theta(x^k) \\ = & p(x^{k+1}) - p(x^k) + \langle x^{k+1} - x^k, \nabla g(x^k) - \xi^k\rangle \\ & + \frac{1}{2}\|x^{k+1} - x^k\|^2_{\mathcal{G}} \\ \le & \langle \nabla g(x^k) + \eta^{k+1} - \xi^k, x^{k+1} - x^k\rangle + \frac{1}{2}\|x^{k+1} - x^k\|^2_{\mathcal{G}} \\ = & -\frac{1}{2}\|x^{k+1} - x^k\|^2_{\mathcal{G}+2\mathcal{T}}. \end{array}$$

This completes the proof of this lemma. □

Now we are ready to prove Theorem 6.

Proof. From the optimality condition at x^{k+1}, we have that

$$0 \in \nabla g(x^k) + (\mathcal{G} + \mathcal{T})(x^{k+1} - x^k) + \partial p(x^{k+1}) - \xi^k.$$

Since $x^{k+1} = x^k$, this implies that

$$0 \in \nabla g(x^k) + \partial p(x^k) - \partial q(x^k),$$

i.e., x^k is a critical point. Observe that the sequence $\{\theta(x^k)\}$ is non-increasing since

$$\theta(x^{k+1}) \le \widehat{\theta}(x^{k+1}; x^k) \le \widehat{\theta}(x^k; x^k) = \theta(x^k), \quad k \ge 0.$$

Suppose that there exists a subsequence $\{x^{k_j}\}$ that converging to $\bar{x}$, i.e., one of the accumulation points of $\{x^k\}$. By Lemma 1 and the assumption that $\mathcal{G} + 2\mathcal{T} \succeq 0$, we know that for all $x \in \mathbb{X}$

$$\begin{array}{l} \widehat{\theta}(x^{k_{j+1}}; x^{k_{j+1}}) = \theta(x^{k_{j+1}}) \\ \le \theta(x^{k_j+1}) \le \widehat{\theta}(x^{k_j+1}; x^{k_j}) \le \widehat{\theta}(x; x^{k_j}). \end{array}$$

By letting $j \to \infty$ in the above inequality, we obtain that

$$\widehat{\theta}(\bar{x}; \bar{x}) \le \widehat{\theta}(x; \bar{x}).$$

By the optimality condition of $\widehat{\theta}(x;\bar{x})$, we have that there exists $\bar{u} \in \partial p(\bar{x})$ and $\bar{v} \in \partial q(\bar{x})$ such that

$$0 \in \nabla g(\bar{x}) + \bar{u} - \bar{v}$$

This implies that $(\nabla g(\bar{x}) + \partial p(\bar{x})) \cap \partial q(\bar{x}) \neq \emptyset$. To establish the rest of this proposition, we obtain from Lemma 1 that

$$\begin{aligned} &\lim_{t\to+\infty} \frac{1}{2}\sum_{i=0}^{t} \|x^{k+1} - x^k\|^2_{\mathcal{G}+2\mathcal{T}} \\ \leq\ &\liminf_{t\to+\infty} \left(\theta(x^0) - \theta(x^{k+1})\right) \leq \theta(x^0) < +\infty\,, \end{aligned}$$

which implies $\lim_{i\to+\infty} \|x^{k+1} - x^i\|_{\mathcal{G}+2\mathcal{T}} = 0$. The proof of this theorem is thus complete by the positive definiteness of the operator $\mathcal{G} + 2\mathcal{T}$. □

8. Discussions on $\mathcal{G}$ and $\mathcal{T}$

Here, we discuss the roles of $\mathcal{G}$ and $\mathcal{T}$. The majorization technique used to handle the smooth function g and the presence of $\mathcal{G}$ are used to make the subproblems (10) in Algorithm (3) more amenable to efficient computations. As can be observed in Theorem 6, the algorithm is convergent if $\mathcal{G} + 2\mathcal{T} \succeq 0$. This indicates that instead of adding the commonly used positive semidefinte or positive definite proximal terms, we allow $\mathcal{T}$ to be indefinite for better practical performance. Indeed, the computational benefit of using indefinite proximal terms has been observed in (Gao & Sun, 2010; Li et al., 2016a). In fact, the introduction of indefinite proximal terms in the DC algorithm is motivated by these numerical evidence. As far as we know, Theorem 6 provides the first rigorous convergence proof of the introduction of the indefinite proximal terms in the DC algorithms. The presence of $\mathcal{G}$ and $\mathcal{T}$ also helps to guarantee the existence of solutions for the subproblems (10). Since $\mathcal{G} + 2\mathcal{T} \succeq 0$ and $\mathcal{G} \succeq 0$, we have that $2\mathcal{G} + 2\mathcal{T} \succeq 0$, i.e., $\mathcal{G} + \mathcal{T} \succeq 0$ (the reverse direction holds when $\mathcal{T} \succeq 0$). Hence, $\mathcal{G} + 2\mathcal{T} \succeq 0$ ($\mathcal{G} + 2\mathcal{T} \succ 0$) implies that subproblems (10) are (strongly) convex problems. Meanwhile, the choices of $\mathcal{G}$ and $\mathcal{T}$ are very much problem dependent. The general principle is that $\mathcal{G} + \mathcal{T}$ should be as small as possible while x^{k+1} is still relatively easy to compute.

References

Bühlmann, Peter and Van De Geer, Sara. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media, 2011.

Gao, Yan and Sun, Defeng. A majorized penalty approach for calibrating rank constrained correlation matrix problems. *technical reprot*, 2010.

Ledoux, Michel and Talagrand, Michel. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer Science & Business Media, 2013.

Li, Min, Sun, Defeng, and Toh, Kim-Chuan. A majorized ADMM with indefinite proximal terms for linearly constrained convex composite optimization. *SIAM Journal on Optimization*, 26(2):922–950, 2016a.

Li, Xudong, Sun, Defeng, and Toh, Kim-Chuan. A Schur complement based semi-proximal ADMM for convex quadratic conic programming and extensions. *Mathematical Programming*, 155(1-2):333–373, 2016b.

Tropp, Joel A. The expected norm of a sum of independent random matrices: An elementary approach. In *High Dimensional Probability VII*, pp. 173–202. Springer, 2016.

Zhang, Anru and Wang, Mengdi. Optimal state compression of Markov processes via empirical low-rank estimation. *arXiv preprint arXiv:1802.02920*, 2018.