Spectral Clustering with a Convex Regularizer on Millions of Images (Proofs)

Maxwell D. Collins Ji Liu Jia Xu
Lopamudra Mukherjee Vikas Sing

September 8, 2014

1 Proofs of Theorems

Theorem 1. *Let V^* be a convergent point of the sequence $\{V_t\}$ generated from equation (5) of the main paper which is in a small ball with radius δ and denote $f(V^*)$ as f^*. Let ϕ be a positive value. If there exists a constant $\delta > 0$ such that $\mathcal{P}_\Omega\left(V_t - \gamma_t(\hat{L}_t V_t + \partial g(V_t))\right)$ is a nonexpansive projection, we have:*

i) *If the stepsize is chosen as $\gamma_t = \frac{\phi\delta}{\sqrt{((M+N)^2+\sigma^2)T}}$ and $\bar{V}_T = (\sum_{t=1}^T \gamma_t)^{-1} \sum_{t=1}^T \gamma_t V_t$, then $\mathbb{E}\left(f(\bar{V}_T)\right) - f^* \leq (\phi + \phi^{-1})\frac{\delta}{2}\Upsilon$.*

ii) *If the step size is chosen as $\gamma_t = \theta_t \frac{f(V_t)-f^*}{(M+N)^2+\sigma^2}$, then $\mathbb{E}(f(\tilde{V}_T)) - f^* \leq \frac{\delta}{\sqrt{\theta_{\min}}}\Upsilon$ where $\tilde{V}_T = \frac{1}{T}\sum_{t=1}^T V_t$, $\theta_t \in (0,2)$ and $\theta_{\min} = \min_t 1 - (\theta_t - 1)^2$.*

Proof. Consider the expansion of $\|V_{t+1} - V^*\|_F^2$:

$$\begin{aligned}
\|V_{t+1} - V^*\|_F^2 &= \|\mathcal{P}_\Omega(V_t - \gamma_t((L+\Delta_t)V_t + \partial g(V_t)) - \mathcal{P}_\Omega(V^*)\|_F^2 \\
&\quad \text{from the local nonexpansive projection property,} \\
&\leq \|V_t - \gamma_t((L+\Delta_t)V_t + \partial g(V_t)) - V^*\|_F^2 \\
&\leq \|V_t - V^*\|^2 + \gamma_t^2 \underbrace{\|(L+\Delta_t)V_t + \partial g(V_t)\|_F^2}_{T_1} \\
&\quad - 2\gamma_t \underbrace{\langle (L+\Delta_t)V_t + \partial g(V_t), V_t - V^*\rangle}_{T_2}.
\end{aligned} \tag{1}$$

Take the conditional expectation of T_1 and T_2 in terms of Δ_t given V_t:

$$\begin{aligned}
\mathbb{E}(T_1) &= \|LV_t + \partial g(V_t)\|_F^2 + \mathbb{E}(\|\Delta_t V_t\|_F^2) + 2\mathbb{E}\langle LV_t + \partial g(V_t), \Delta_t V_t\rangle \\
&= \mathbb{E}(\|LV_t + \partial g(V_t)\|_F^2) + \mathbb{E}(\|\Delta_t V_t\|_F^2) \\
&\leq (M+N)^2 + \sigma^2
\end{aligned} \tag{2}$$

$$\mathbb{E}(T_2) = \mathbb{E}\langle LV_t + \partial g(V_t), V_t - V^*\rangle \geq \mathbb{E}(f(V_t)) - f^*. \tag{3}$$

Take the expectation of both sides of (1) in terms of all random variables, together with (2), and (3), we have

$$2\gamma_t(\mathbb{E}(f(V_t)) - f^*) \leq \mathbb{E}\|V_t - V^*\|_F^2 - \mathbb{E}(\|V_{t+1} - V^*\|_F^2) + \gamma_t^2((M+N)^2 + \sigma^2) \tag{4}$$

which implies that

$$\begin{aligned}
2\sum_{t=1}^{T}\gamma_t(\mathbb{E}(f(V_t)) - f^*) &\leq \mathbb{E}\|V_1 - V^*\|_F^2 + ((M+N)^2 + \sigma^2)\sum_{t=1}^{T}\gamma_t^2 \\
&\leq \delta^2 + ((M+N)^2 + \sigma^2)\sum_{t=1}^{T}\gamma_t^2.
\end{aligned}$$

Also note that

$$\sum_{t=1}^{T}\gamma_t = \frac{\phi\delta\sqrt{T}}{\sqrt{(M+N)^2+\sigma^2}} \qquad \sum_{t=1}^{T}\gamma_t^2 = \frac{(\phi\delta)^2}{(M+N)^2+\sigma^2},$$

and

$$\frac{\sum_{t=1}^{T}\gamma_t\mathbb{E}(f(v_t))}{\sum_{t=1}^{T}\gamma_t} = \frac{\mathbb{E}\sum_{t=1}^{T}\gamma_t f(v_t)}{\sum_{t=1}^{T}\gamma_t} \leq \mathbb{E}f(\bar{V}_t). \quad \text{(from the convexity of } f(V_t))$$

It follows that

$$\begin{aligned}
\frac{\sum_{t=1}^{T}\gamma_t(\mathbb{E}(f(V_t)) - f^*)}{\sum_{t=1}^{T}\gamma_t} &\leq \frac{\delta^2 + ((M+N)^2+\sigma^2)\sum_{t=1}^{T}\gamma_t^2}{2\sum_{t=1}^{T}\gamma_t} \\
\Rightarrow \ \mathbb{E}(f(\bar{V}_t) - f^*) &\leq \frac{\delta^2 + ((M+N)^2+\sigma^2)\sum_{t=1}^{T}\gamma_t^2}{2\sum_{t=1}^{T}\gamma_t} \\
&= \frac{\delta^2 + ((M+N)^2+\sigma^2)\frac{(\phi\delta)^2}{(M+N)^2+\sigma^2}}{2\frac{\phi\delta\sqrt{T}}{\sqrt{(M+N)^2+\sigma^2}}} \\
&= (\phi + \phi^{-1})\frac{\delta\sqrt{(M+N)^2+\sigma^2}}{2\sqrt{T}}
\end{aligned}$$

proving the first claim. Next we prove the second claim. From (1), (2), and (3), we have

$$\begin{aligned}
\mathbb{E}(\|V_{t+1} - V^*\|_F^2) &\leq \|V_t - V^*\|_F^2 + \gamma_t^2((M+N)^2+\sigma^2) - 2\gamma_t(f(V_t) - f^*) \\
&\leq \|V_t - V^*\|_F^2 - \frac{(f(V_t)-f^*)^2}{(M+N)^2+\sigma^2} + ((M+N)^2+\sigma^2)\left(\gamma_t - \frac{f(V_t)-f^*}{(M+N)^2+\sigma^2}\right)^2 \\
&\leq \|V_t - V^*\|_F^2 - \frac{(1-(1-\theta_t)^2)(f(V_t)-f^*)^2}{(M+N)^2+\sigma^2} \\
&\leq \|V_t - V^*\|_F^2 - \frac{\theta_{min}(f(V_t)-f^*)^2}{(M+N)^2+\sigma^2}.
\end{aligned}$$

It follows that

$$\frac{\theta_{min}}{(M+N)^2+\sigma^2}\mathbb{E}(f(V_t)-f^*)^2 \le \mathbb{E}(\|V_t - V^*\|_F^2) - \mathbb{E}(\|V_{t+1}-V^*\|_F^2). \tag{5}$$

Taking $t = 0, 1, \cdots, T-1$ in (5) respectively and summarizing all of them, we obtain

$$\begin{aligned} \frac{\theta_{min}}{(M+N)^2+\sigma^2}\sum_{t=1}^{T}\mathbb{E}(f(V_t)-f^*)^2 &\le \mathbb{E}(\|V_1 - V^*\|_F^2) \le \delta^2 \\ \Rightarrow\ T^{-1}\sum_{t=1}^{T}\mathbb{E}(f(V_t)-f^*)^2 &\le \frac{\delta^2((M+N)^2+\sigma^2)}{T\theta_{min}}. \end{aligned}$$

Together with

$$\begin{aligned} T^{-1}\sum_{t=1}^{T}\mathbb{E}(f(V_t)-f^*)^2 &\ge T^{-1}\sum_{t=1}^{T}(\mathbb{E}(f(V_t))-f^*)^2 \\ &\ge (T^{-1}\sum_{t=1}^{T}\mathbb{E}(f(V_t))-f^*)^2 \ge (\mathbb{E}(f(\tilde{V}_T))-f^*)^2. \end{aligned}$$

The last inequality uses Jensen's inequality, that is, $\mathbb{E}f(x) \ge f(\mathbb{E}(x))$ holds for any convex function. We prove the second claim. □

Denote $[t]$ as a subset of coordinates of $V \in \mathbb{R}^{n\times p}$, which is randomly selected at iteration t. To make our following discussion simpler, we assume that the size of $[t]$ is a constant and denote the ratio $R := \frac{np}{|[t]|}$. Consider the following update for V_{t+1}, also appearing in equation (9) of the main paper:

$$V_{t+1} = \mathcal{P}_\Omega(V_t - \gamma_t \partial_{[t]} f(V_t)) \tag{6}$$

Theorem 2. *Let V^* be a convergent point of the sequence $\{V_t\}$ generated from (6) which is in a small ball with radius δ and denote $f(V^*)$ as f^*. Let ϕ be a positive value. Let $\bar{\Upsilon} := \frac{(M+N)R}{\sqrt{T}}$. If there exists a constant $\delta > 0$ such that $\mathcal{P}_\Omega\left(V_t - \gamma_t\partial_{[t]}f(V_t)\right)$ is a nonexpansive projection, we have:*

i) *If the stepsize is chosen as $\gamma_t = \frac{\phi\delta}{(M+N)\sqrt{T}}$ and $\bar{V}_T = (\sum_{t=1}^T \gamma_t)^{-1}\sum_{t=1}^T \gamma_t V_t$, then $\mathbb{E}\left(f(\bar{V}_T)\right) - f^* \le (\phi+\phi^{-1})\frac{\delta}{2}\bar{\Upsilon}$.*

ii) *If the step size is chosen as $\gamma_t = \theta_t\frac{f(V_t)-f^*}{R(M+N)^2}$, then $\mathbb{E}(f(\tilde{V}_T)) - f^* \le \frac{\delta}{\sqrt{\theta_{\min}}}\bar{\Upsilon}$ where $\tilde{V}_T = \frac{1}{T}\sum_{t=1}^T V_t$, $\theta_t \in (0,2)$ and $\theta_{\min} = \min_t 1-(\theta_t-1)^2$.*

This theorem basically shows the convergence rate for (6) is $O(1/\sqrt{T})$, which is the same as the full projection in (5) of the main paper. The speedup property is also similar: both convergence rates are proportional to R. R is basically the inverse of the block size of $[t]$. Hence, when the block size increases x times, the required iterations to achieve the given accuracy decreases x times.

Proof. Consider the expansion of $\|V_{t+1} - V^*\|_F^2$:

$$\begin{aligned}
&\|V_{t+1} - V^*\|_F^2 = \|\mathcal{P}_\Omega(V_t - \gamma_t \hat{\partial}_{[t]} f(V_t)) - \mathcal{P}_\Omega(V^*)\|_F^2 \\
&\quad \leq \|V_t - \gamma_t \hat{\partial}_{[t]} f(V_t) - V^*\|_F^2 \quad \text{(from the local nonexpansive projection property)} \\
&\quad \leq \|V_t - V^*\|^2 + \gamma_t^2 \underbrace{\|\hat{\partial}_{[t]} f(V_t)\|_F^2}_{T_3} - 2\gamma_t \underbrace{\langle \hat{\partial}_{[t]} f(V_t), V_t - V^* \rangle}_{T_4}.
\end{aligned} \tag{7}$$

Take the conditional expectation of T_1 and T_2 in terms of Δ_t given V_t:

$$\mathbb{E}(T_3) = \mathbb{E}(\|\partial_{[t]} f(V_t)\|_F^2) \leq \mathbb{E}\|\partial f(V_t)\|_F^2 = \mathbb{E}\|LV_t + \partial g(V_t)\|_F^2 \leq (M+N)^2 \tag{8}$$

$$\mathbb{E}(T_4) = \mathbb{E}\langle \partial_{[t]} f(V_t), V_t - V^* \rangle = \frac{1}{R}\mathbb{E}\langle \partial f(V_t), V_t - V^* \rangle \geq \frac{1}{R}(\mathbb{E}(f(V_t)) - f^*). \tag{9}$$

Take the expectation on both sides of (7) in terms of all random variables, we have

$$2\gamma_t(\frac{1}{R}(\mathbb{E}f(V_t) - f^*)) \leq \mathbb{E}\|V_t - V^*\|^2 - \mathbb{E}\|V_{t+1} - V^*\|_F^2 + \gamma_t^2 (M+N)^2.$$

The rest of the proof can follow the proof of Theorem 1 by simply treating "$\frac{1}{R}(\mathbb{E}f(V_t) - f^*)$" as "$\mathbb{E}f(V_t) - f^*$" in (4). $\square$