Orthogonal Constraint Spherical Matrix Factorization

Anonymous*

1 Convergence Analysis

Theorem 1 (Subsequence convergence). *Let $\{W_k\}_{k\geq 0} = \{(U_k, V_k)\}_{k\geq 0}$ be the sequence generated by Algorithm 1 with constant step size $\lambda, \mu > L_c$. Then the sequence $\{W_k\}_{k\geq 0}$ is bounded and obeys the following properties:*

(P1) sufficient decrease:

$$f(W_{k-1}) - f(W_k) \geq \frac{\min(\lambda,\mu) - L_c}{2}\|W_k - W_{k-1}\|_F^2 \tag{1}$$

implying

$$\lim_{k\to\infty} \|W^{k-1} - W^k\|_F = 0. \tag{2}$$

(P2) the sequence $\{f(W_k)\}$ converges to some $\overline{f} \geq 0$.

(P3) denote $\mathbb{C}(W_0)$ (depending on W_0) as the set of all limit points of the iterates $\{W_k\}$. Then all the limit points $W^\star$ are critical points of f and have the same function value

$$f(W^\star) = \overline{f}. \tag{3}$$

Further, $\mathbb{C}(W_0)$ is a nonempty, compact and connected set and satisfies

$$\lim_{k\to\infty} \operatorname{dist}(W_k, \mathbb{C}(W_0)) = 0 \tag{4}$$

Proof of Theorem 1. Before proving Theorem 1, we give out some necessary definition.

Definition 1. *[Attouch* et al.*, 2013] Let $f : \mathbb{R}^d \to (-\infty, \infty]$ be a proper and lower semi-continuous function, whose domain is defined as*

$$\operatorname{dom} f := \{u \in \mathbb{R}^n : f(u) < \infty\}.$$

The (Fréchet) subdifferential ∂f of f at u is defined by

$$\partial f(u) = \left\{ z : \lim_{v\to u} \inf \frac{f(v) - f(u) - \langle z, v-u\rangle}{\|u - v\|} \geq 0 \right\}$$

for any $u \in \operatorname{dom} h$ and $\partial f(u) = \emptyset$ if $u \notin \operatorname{dom} f$.

We say u is a limiting critical point, or simply a critical point of f if $0 \in \partial f(u)$.

*This is the supplementary of Paper ID: 5740.

We now turn to prove Theorem 1.

(P1): First note that for all k, according to our alternating minimization method, we always have $\delta_{\mathbb{U}}(U_k) = \delta_{\mathbb{V}}(V_k) = 0$ and thus $f(W_k) = h(W_k)$.

Since $h(U, V)$ has Lipschitz continuous gradient at $U \in \mathbb{U}, V \in \mathbb{V}$ with Lipschitz gradient L_c and $\lambda > L_c$, we define $h_{L_c}(U, U', V)$ as proximal regularization of $h(U, V)$ linearized at U', V:

$$h(U', V) + \langle \nabla_U h(U', V), U - U'\rangle + \frac{L_c}{2}\|U - U'\|_F^2,$$

By the definition of Lipschitz continuous gradient and Taylor expansion, we have

$$h(U, V) \leq h_{L_c}(U, U', V). \tag{5}$$

Also by the definition of proximal map, we get:

$$\begin{aligned} U_k = \arg\min_U \delta_{\mathbb{U}}(U) + \frac{\mu}{2}\|U - U_{k-1}\|_F^2 \\ + \langle \nabla_U h(U_{k-1}, V_{k-1}), U - U_{k-1}\rangle \end{aligned} \tag{6}$$

and hence we take $U_k = U$, which implies that

$$\begin{aligned} &\delta_{\mathbb{U}}(U_k) + \frac{\mu}{2}\|U_k - U_{k-1}\|_F^2 \\ &+ \langle \nabla_U h(U_{k-1}, V_{k-1}), U_k - U_{k-1}\rangle \leq \delta_{\mathbb{U}}(U_{k-1}) \end{aligned} \tag{7}$$

Combining Eq. (5) to Eq. (7), we have:

$$\begin{aligned} &h(U_k, V_{k-1}) + \delta_{\mathbb{U}}(U_k) \\ &\leq h(U_{k-1}, V_{k-1}) + \langle \nabla_U h(U_{k-1}, V_{k-1}), U_k - U_{k-1}\rangle \\ &+ \frac{L_c}{2}\|U_k - U_{k-1}\|_F^2 + \delta_{\mathbb{U}}(U_k) \\ &\leq h(U_{k-1}, V_{k-1}) + \frac{L_c}{2}\|U_k - U_{k-1}\|_F^2 \\ &+ \delta_{\mathbb{U}}(U_{k-1}) - \frac{\mu}{2}\|U_k - U_{k-1}\|_F^2 \\ &= h(U_{k-1}, V_{k-1}) + \delta_{\mathbb{U}}(U_{k-1}) - \frac{\mu - L_c}{2}\|U_k - U_{k-1}\|_F^2, \end{aligned} \tag{8}$$

Similarly, we have

$$\begin{aligned} &h(U_k, V_k) - h(U_k, V_{k-1}) + \delta_{\mathbb{V}}(V_k) - \delta_{\mathbb{V}}(V_{k-1}) \\ &\leq -\frac{\lambda - L_c}{2}\|V_k - V_{k-1}\|_F^2 \end{aligned} \tag{9}$$

which together with the above equation gives Eq. (1). Now repeating Eq. (1) for all k will give

$$(\min(\lambda, \mu) - L_c) \sum_{k=1}^{\infty} \|W_k - W_{k-1}\|_F^2 \le 2f(W_0), \quad (10)$$

which gives Eq. (2).

Remark 1. *In our proposed algorithm, since in every update, our solution is closed while satisfying the constraints, thus in fact $\delta_{\mathbb{U}}$ and $\delta_{\mathbb{V}}$ are 0, and ∞ is never achieved.*

(P2) It follows from Eq. (16) that $\{f(W_k)\}_{k\ge 0}$ is a decreasing sequence. Due to the fact that f is lower bounded as $f(W_k) \ge 0$ for all k, we conclude that $\{f(W_k)\}_{k\ge 0}$ is convergent to some constant $\overline{f} \ge 0$.

(P3) Extract any convergent subsequence $\{W_{k'} = (U_{k'}, V_{K'})\}$ from $\{W_k\}$ and denote the limit point of this subsequence as $W^\star$. Since $U_{k'} \in \mathbb{U}, V_{k'} \in \mathbb{V}$ for all k' and both of the sets $\mathbb{U}$ and $\mathbb{V}$ are closed, we have $U^\star \in \mathbb{U}, V^\star \in \mathbb{V}$. Since h is continuous, we have

$$\begin{aligned}\lim_{k'\to\infty} f(W_{k'}) &= \lim_{k'\to\infty} h(U_{k'}, V_{k'}) + \delta_{\mathbb{U}}(U_{k'}) + \delta_{\mathbb{V}}(V_{k'}) \\ &= f(W^\star), \end{aligned} \quad (11)$$

which together with the fact that $\overline{f} = \lim_{k\to\infty} f(W_{k'})$ gives Eq. (3).

To show $W^\star$ is a critical point, we first consider Eq. (6) and the optimality condition yields:

$$\nabla_U h(U_{k-1}, V_{k-1}) + \mu(U_k - U_{k-1}) + \partial\delta_{\mathbb{U}}(U_k) = 0. \quad (12)$$

Similarly, we have

$$\nabla_V h(U_k, V_{k-1}) + \lambda(V_k - V_{k-1}) + \partial\delta_{\mathbb{V}}(V_k) = 0. \quad (13)$$

Now, define

$$\underbrace{\nabla_U h(U_k, V_k) + \partial\delta_{\mathbb{U}}(U_k)}_{A_k} \quad \text{and} \quad \underbrace{\nabla_V h(U_k, V_k) + \partial\delta_{\mathbb{V}}(V_k)}_{B_k}.$$

Thus, we have

$$A_k \in \partial_U f(U_k, V_k), B_k \in \partial_V f(U_k, V_k). \quad (14)$$

It follows from the above that

$$\begin{aligned} &\lim_{k\to\infty} \|A_k\|_F \\ &\le \lim_{k\to\infty} \|\nabla_U h(U_k, V_k) - \nabla_U h(U_{k-1}, V_{k-1})\|_F + \mu\|U_k - U_{k-1}\|_F \\ &\le \lim_{k\to\infty} (L_c + \mu)\|W_k - W_{k-1}\|_F = 0. \end{aligned} \quad (15)$$

Similarly, we have

$$\lim_{k\to\infty} \|B_k\|_F \le \lim_{k\to\infty} (L_c + \lambda)\|W_k - W_{k-1}\|_F = 0. \quad (16)$$

Then we have:

$$\operatorname{dist}(0, \partial f(W_k)) \le \mathcal{L}_g \|W_k - W_{k-1}\|_F \quad (17)$$

where $\mathcal{L}_g := (2L_c + \mu + \lambda)$. Owing to the closedness properties of $\partial f(W_{k'})$, we finally obtain $0 \in \partial f(W^\star)$. Thus, $W^\star$ is a critical point of f.

The remaining proof regarding to the properties of $\mathbb{C}(W_0)$ follows from [Bolte *et al.*, 2014] by using the regularity of $\{W_k\}$ (assertion (P1) of Theorem 1). □

Theorem 2 (Sequence convergence)**.** *The sequence $\{W_k\}_{k\ge 0}$ generated by Algorithm 1 with a constant step size $\lambda, \mu > L_c$ is global-sequence convergence.*

Remark 2. *Theorem 2 is much stronger than Theorem 1, since we are not guaranteed that the iterates $\{W_k\}$ generated by Algorithm 1 would converge to a limit point only by Theorem 1. Theorem 2 fulfills this gap by directly showing the iterates $\{W_k\}$ generated by Algorithm 1 converges to a critical point, and as an consequence, the set of limit point $\mathbb{C}(W_0)$ becomes a singleton.*

Before proving Theorem 2, we give out another important definition.

Definition 2 (Kurdyka-Lojasiewicz (KL) property). *[Bolte* et al.*, 2007] We say a proper semi-continuous function $h(\boldsymbol{u})$ satisfies Kurdyka-Lojasiewicz (KL) property, if $\overline{\boldsymbol{u}}$ is a critical point of $h(\boldsymbol{u})$, then there exist $\delta > 0,\ \theta \in [0, 1),\ C_1 > 0,\ s.t.$*

$$|h(\boldsymbol{u}) - h(\overline{\boldsymbol{u}})|^\theta \le C_1 \operatorname{dist}(0, \partial h(\boldsymbol{u})), \quad \forall\ \boldsymbol{u} \in B(\overline{\boldsymbol{u}}, \delta)$$

We mention that the above KL property(also known as KL inequality) states the regularity of $h(u)$ around its critical point u and the KL inequality trivially holds at non-critical point. There are a very large set of functions satisfying the KL inequality including any semi-algebraic functions [Attouch *et al.*, 2013; Bolte *et al.*, 2014]. Clearly, the objective function f is semi-algebraic as both h, $\delta_{\mathbb{U}}$ and $\delta_{\mathbb{V}}$ are semi-algebraic [Bolte *et al.*, 2014].

Lemma 1 (Uniform KL property)**.** *There exist $\delta_0 > 0,\ \theta_{KL} \in [0, 1),\ C_{KL} > 0$ such that for all W s.t. $\operatorname{dist}((W), \mathbb{C}(W_0)) \le \delta_0$:*

$$\left|f(W) - \overline{f}\right|^{\theta_{KL}} \le C_{KL} \operatorname{dist}(0, \partial f(W)) \quad (18)$$

with $\overline{f}$ denoting the limiting function value defined in P (2) of Theorem 1.

Proof. First we recognize the union $\bigcup_i B(W_i^\star, \delta_i)$ forms an open cover of $\mathbb{C}(W_0)$ with $W_i^\star$ representing all points in $\mathbb{C}(W_0)$ and δ_i to be chosen so that the the following KL property of f at $W_i^\star \in \mathbb{C}(W_0)$ holds:

$$\left|f(W) - \overline{f}\right|^{\theta_i} \le C_i \operatorname{dist}(0, \partial f(W)) \ \ \forall\ (W) \in B(W_i^\star, \delta_i)$$

where we have used all $f(W_i^\star) = \overline{f}$ by assertion (P3) of Theorem 1. Then due to the compactness of the set $\mathbb{C}(W_0)$, it has a finite subcover $\bigcup_{i=1}^p B(W_{k_i}^\star, \delta_{k_i})$ for some positive integer p. Now combining all, we have for all $W \in \bigcup_{i=1}^p B(W_{k_i}^\star, \delta_{k_i})$,

$$\left|f(W) - \overline{f}\right|^{\theta_{KL}} \le C_{KL} \operatorname{dist}(0, \partial f(W)) \quad (19)$$

with $\theta_{KL} = \max_{i=1}^p \{\theta_{k_i}\}$ and $C_{KL} = \max_{i=1}^p \{C_{k_i}\}$. Finally, since $\bigcup_{i=1}^p B(W_{k_i}^\star, \delta_{k_i})$ is an open cover of $\mathbb{C}(W_0)$, there exists a sufficiently small number δ_0 so that

$$\{(W) : \operatorname{dist}(W, \mathbb{C}(W_0)) \le \delta_0\} \subset \bigcup_{i=1}^p B(W_i^\star, \delta_{k_i}).$$

Therefore, Eq. (19) holds whenever $\operatorname{dist}(W, \mathbb{C}(W_0)) \le \delta_0$. □

Now we are ready to prove Theorem 2.

Proof of Theorem 2. First of all, following from assertion (P3) in Theorem 1, there exits a positive integer k_0 so that $\mathrm{dist}(W_k, \mathbb{C}(W_0)) \le \delta_0$ for all $k \ge k_0$. Now using Lemma 1, we have

$$[f(W_k) - f(W^\star)]^{\theta_{KL}} \le C_{KL}\,\mathrm{dist}(0, \partial f(W_k)), \quad \forall k \ge k_0. \tag{20}$$

In the subsequent analysis, we restrict to $k \ge k_0$. Construct a concave function $x^{1-\theta}$ for some $\theta \in [0, 1)$ with domain $x > 0$. Obviously, by the concavity, we have

$$x_2^{1-\theta} - x_1^{1-\theta} \ge (1-\theta)x_2^{-\theta}(x_2 - x_1), \forall x_1 > 0, x_2 > 0$$

Replacing x_1 by $f(W_{k+1}) - f(W^\star)$ and x_2 by $f(W_k) - f(W^\star)$, choosing $\theta = \theta_{KL}$ and using the sufficient decrease property, we have

$$\begin{aligned}
&[f(W_k) - f(W^\star)]^{1-\theta_{KL}} - [f(W_{k+1}) - f(W^\star)]^{1-\theta_{KL}} \\
&\ge (1-\theta_{KL})\frac{f(W_k) - f(W_{k+1})}{[f(W_k) - f(W^\star)]^{\theta_{KL}}} \\
&\ge \frac{\lambda(1-\theta_{KL})}{2C_{KL}}\frac{\|W_k - W_{k+1}\|_F^2}{\mathrm{dist}(0, \partial f(W_k))} \\
&\ge \frac{\lambda(1-\theta_{KL})}{2C_{KL}\mathcal{L}_g}\frac{\|W_k - W_{k+1}\|_F^2}{\|W_k - W_{k-1}\|_F} \\
&= \kappa(\frac{\|W_k - W_{k+1}\|_F^2}{\|W_k - W_{k-1}\|_F} + \|W_k - W_{k-1}\|_F) - \kappa\|W_k - W_{k-1}\|_F \\
&\ge \kappa\left(2\|W_k - W_{k+1}\|_F - \|W_k - W_{k-1}\|_F\right)
\end{aligned}$$

where we have used Eq. (20) in the third line and Eq. (17) in the fourth line. And accordingly, we have:

$$\begin{aligned}
&2\|W_k - W_{k+1}\|_F - \|W_k - W_{k-1}\|_F \\
&\le \beta\left([f(W_k) - f(W^\star)]^{1-\theta} - [f(W_{k+1}) - f(W^\star)]^{1-\theta_{KL}}\right)
\end{aligned} \tag{21}$$

with $\kappa := \frac{\lambda(1-\theta_{KL})}{2C_{KL}\mathcal{L}_g}$ and $\beta := \left(\frac{\lambda(1-\theta_{KL})}{2C_{KL}\mathcal{L}_g}\right)^{-1}$.

Summing the above inequalities up from some $\widetilde{k} > k_0$ to infinity yields

$$\begin{aligned}
&\sum_{k=\widetilde{k}}^{\infty}\|W_k - W_{k+1}\|_F \\
&\le \|W_{\widetilde{k}} - W_{\widetilde{k}-1}\|_F + \beta[f(W_{\widetilde{k}}) - f(W^\star)]^{1-\theta_{KL}}
\end{aligned} \tag{22}$$

implying $\sum_{k=\widetilde{k}}^{\infty}\|W_k - W_{k+1}\|_F < \infty$. Following some standard arguments one can see that

$$\limsup_{t\to\infty, t_1, t_2 \ge t}\|W_{t_1} - W_{t_2}\|_F = 0$$

which implies that the sequence $\{W_k\}$ is Cauchy, and hence convergent. Hence, the limit point set $\mathcal{C}(W_0)$ is singleton $W^\star$. □

Theorem 3 (Convergence Rate)**.** *The convergence rate is at least sub-linear.*

Towards that end, we first know from the above argument that $\{W_k\}$ converges to some point $W^\star$, i.e., $\lim_{k\to\infty} W^k = W^\star$. Then using Eq. (22) and the triangle inequality, we obtain

$$\begin{aligned}
\|W_{\widetilde{k}} - W^\star\|_F &\le \sum_{k=\widetilde{k}}^{\infty}\|W_k - W_{k+1}\|_F \\
&\le \|W_{\widetilde{k}} - W_{\widetilde{k}-1}\|_F + \beta[f(W_{\widetilde{k}}) - f(W^\star)]^{1-\theta_{KL}}
\end{aligned} \tag{23}$$

which indicates the convergence rate of $W_{\widetilde{k}} \to W^\star$ is at least as fast as the rate that $\|W_{\widetilde{k}} - W_{\widetilde{k}-1}\|_F + \beta[f(W_{\widetilde{k}}) - f(W^\star)]^{1-\theta_{KL}}$ converges to 0. In particular, the second term $\beta[f(W_{\widetilde{k}}) - f(W^\star)]^{1-\theta_{KL}}$ can be controlled:

$$\begin{aligned}
&\beta[f(W_{\widetilde{k}}) - f(W^\star)]^{\theta_{KL}} \le \beta C_{KL}\,\mathrm{dist}(0, \partial f(W_{\widetilde{k}})) \\
&\le \underbrace{\beta C_{KL}(2B_0 + \lambda + \|\boldsymbol{X}\|_F)}_{:=\alpha}\|W_{\widetilde{k}} - W_{\widetilde{k}-1}\|_F
\end{aligned} \tag{24}$$

Plugging Eq. (24) back to Eq. (23), we then have

$$\sum_{k=\widetilde{k}}^{\infty}\|W_k - W_{k+1}\|_F \le \|W_{\widetilde{k}} - W_{\widetilde{k}-1}\|_F + \alpha\|W_{\widetilde{k}} - W_{\widetilde{k}-1}\|_F^{\frac{1-\theta_{KL}}{\theta_{KL}}}.$$

We divide the following analysis into two cases based on the value of the KL exponent θ_{KL}.

Case I : $\theta_{KL} \in [0, \frac{1}{2}]$. This case means $\frac{1-\theta_{KL}}{\theta_{KL}} \ge 1$. We define $P_{\widetilde{k}} = \sum_{i=\widetilde{k}}^{\infty}\|W_{i+1} - W_i\|_F$,

$$P_{\widetilde{k}} \le P_{\widetilde{k}-1} - P_{\widetilde{k}} + \alpha\left[P_{\widetilde{k}-1} - P_{\widetilde{k}}\right]^{\frac{1-\theta_{KL}}{\theta_{KL}}}. \tag{25}$$

Since $P_{\widetilde{k}-1} - P_{\widetilde{k}} \to 0$, there exists a positive integer k_1 such that $P_{\widetilde{k}-1} - P_{\widetilde{k}} < 1$, $\forall\, \widetilde{k} \ge k_1$. Thus,

$$P_{\widetilde{k}} \le (1+\alpha)\left(P_{\widetilde{k}-1} - P_{\widetilde{k}}\right), \quad \forall\, \widetilde{k} \ge \max\{k_0, k_1\},$$

which implies that

$$P_{\widetilde{k}} \le \rho \cdot P_{\widetilde{k}-1}, \quad \forall\, \widetilde{k} \ge \max\{k_0, k_1\}, \tag{26}$$

where $\rho = \frac{1+\alpha}{2+\alpha} \in (0, 1)$. This together with Eq. (23) gives the linear convergence rate

$$\|W_k - W^\star\|_F \le \mathcal{O}(\rho^{k-\overline{k}}),\ \forall\ k \ge \overline{k}. \tag{27}$$

where $\overline{k} = \max\{k_0, k_1\}$.

Case II : $\theta_{KL} \in (1/2, 1)$. This case means $\frac{1-\theta_{KL}}{\theta_{KL}} \le 1$. Based on the former results, we have

$$P_{\widetilde{k}} \le (1+\alpha)\left[P_{\widetilde{k}-1} - P_{\widetilde{k}}\right]^{\frac{1-\theta_{KL}}{\theta_{KL}}},\ \forall\, \widetilde{k} \ge \max\{k_0, k_1\}$$

which gives

$$P_{\widetilde{k}}^{\frac{1-2\theta_{KL}}{1-\theta_{KL}}} - P_{\widetilde{k}-1}^{\frac{1-2\theta_{KL}}{1-\theta_{KL}}} \ge \zeta,\ \forall\ k \ge \overline{k}$$

for some $\zeta > 0$. Then repeating and summing up the above inequality from $\overline{k} = \max\{k_0, k_1\}$ to any $k > \overline{k}$, we can conclude

$$P_{\widetilde{k}} \le \left[P_{\widetilde{k}-1}^{\frac{1-2\theta_{KL}}{1-\theta_{KL}}} + \zeta(\widetilde{k} - \overline{k})\right]^{-\frac{1-\theta_{KL}}{2\theta_{KL}-1}} = \mathcal{O}((\widetilde{k} - \overline{k})^{-\frac{1-\theta_{KL}}{2\theta_{KL}-1}})$$

Finally, the following sublinear convergence holds

$$\|W_k - W^\star\|_F \le \mathcal{O}((k - \overline{k})^{-\frac{1-\theta_{KL}}{2\theta_{KL}-1}}),\ \forall\ k > \overline{k}.$$

We end this proof by commenting that both linear and sublinear convergence rate are closely related to the KL exponent θ_{KL} at the critical point $W^\star$.

References

[Attouch and Bolte, 2009] Hedy Attouch and Jérôme Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming*, 116(1-2):5–16, 2009.

[Attouch *et al.*, 2013] Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Mathematical Programming*, 137(1-2):91–129, 2013.

[Bolte *et al.*, 2007] Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. The łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007.

[Bolte *et al.*, 2014] Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization or nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1-2):459–494, 2014.