TR Local Convergence 0607

Jiaxin Hu

06/07/2020

I am not sure whether expression is the same as " module the orthogonal transformation " in our last version

LOCAL CONVERGENCE PROPERTY FOR TR ALGORITHM 1

Here we study the local convergence property of iterates generated by Algorithm 1.

Theorem 1 (Local Convergence)**.** *Assume the solution to each block update in the alternating optimization exists and is unique. Let $\mathcal{B}^* = (\mathcal{C}^*, \{M_k^*\})$ be a local minimizer of $\mathcal{L}$ and assume the Hessian at $\mathcal{B}^*$ is strictly negative definite in every direction except those tangent to the orthogonal transformation of M_k^*. Then the sequence $\mathcal{B}^{(t)} = \mathcal{C}^{(t)} \times \{M_k^{(t)}\}$ generated by alternating algorithm linearly converges to $\mathcal{B}^*$; i.e.*

$$\left\|\mathcal{B}^{(t)} - \mathcal{B}^*\right\|_F \leq \rho^t(\left\|\mathcal{C}^{(0)} - \mathcal{C}\right\|_F + \sum_{k=1}^{K} \left\|M_k^{(0)} - M_K^*\right\|_F),$$

for any initialization $(\mathcal{C}^{(0)}, \{M_k^{(0)}\})$ sufficiently close to $(\mathcal{C}^*, \{M_k^*\})$. Here $t \in \mathbb{N}^+$ is the iteration number and $\rho \in (0,1)$ is a contraction parameter.

PROOF

Let $S : \mathbb{R}^d \mapsto \mathbb{R}^d$ denote the update mapping that sends t-th iterate to $(t+1)$-th iterate, where $d = r_1 \dots r_K + \sum_k r_k(d_k - 1)$ is the number of decision variables. Then, $S(\mathcal{A}^{(t)}) = \mathcal{A}^{(t+1)}$ and $S(\mathcal{A}^*) = \mathcal{A}^*$.

According to the alternating algorithm, there are $K+1$ micro-steps for each block of decision variables in one iteration. That implies S is composed by $K+1$ block-wise mappings. Next we prove S is continuously differentiable through decomposing the S.

To decompose S, let $C_k : \mathbb{R}^{d - r_k(d_k-1)} \mapsto \mathbb{R}^{r_k(d_k-1)}$ denote the mapping to obtain M_k given $(\mathcal{C}, M_1, \dots, M_{k-1}, M_{k+1}, \dots, M_K)$, for $\forall k \in [K]$ and let $C_{K+1} : \mathbb{R}^{d - r_1 \dots r_K} \mapsto \mathbb{R}^{r_1 \dots r_K}$ denote the mapping to obtain $\mathcal{C}$ given $\{M_k\}$:

$$C_k(\mathcal{C}, M_1, \dots, M_{k-1}, M_{k+1}, \dots, M_K) \triangleq C_k, \text{ where } \nabla_{M_k}\mathcal{L}(\mathcal{C}, M_1, \dots, M_{k-1}, C_k, M_{k+1}, \dots, M_K) = 0 \quad (1)$$
$$\text{and } C_{K+1}(\{M_k\}) \triangleq C_{K+1}, \text{ where } \nabla_{\mathcal{C}}\mathcal{L}(C_{K+1}, \{M_k\}) = 0.$$

Because each block update exists a unique solution, there exists such a C_k satisfies the condition 1 and $\nabla_{M_k, M_k}\mathcal{L}(\mathcal{C}, M_1, \dots, M_{k-1}, C_k, M_{k+1}, \dots, M_K)$ is non-singular $\forall k \in [K]$. By implicit function theorem, $C_k, \forall k \in [K]$ is continuously differentiable. Similarly, C_{K+1} is also continuously differentiable.

Then we define the block-wise mapping $S_k : \mathbb{R}^d \mapsto \mathbb{R}^d$ based on C_k:

$$\begin{aligned} &S_k(\mathcal{C}, \{M_k\}) \triangleq (\mathcal{C}, M_1, \dots, M_{k-1}, C_k, M_{k+1}, \dots, M_K), \forall k \in [K] \\ &S_{K+1}(\mathcal{C}, \{M_k\}) \triangleq (C_{K+1}, \{M_k\}) \end{aligned}$$

Since C_ks are continuously differentiable, $S_k, \forall k \in [K+1]$ are continuously differentiable. The update mapping S can be decomposed as:

$$S(\mathcal{C}^{(t)}, \{M_k^{(t)}\}) = S_{K+1} \circ \cdots \circ S_1(\mathcal{C}^{(t)}, \{M_k^t\}).$$

Therefore S is continuously differentiable.

Next, we want to find the first order derivative of S at $(\mathcal{C}^*, \{M_k^*\})$. For simplicity, let $\mathcal{A} = (\mathcal{C}, \{M_k\})$ denote the decision variables. Define the function $F_k : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}^{r_k(d_k-1)}$ for $\forall k \in [K]$ as:

$$F_k(\mathcal{A}, \mathcal{A}') \triangleq \nabla_{M_k} \mathcal{L}(\mathcal{C}', M_1, \dots, M_k, M'_{k+1}, \dots, M'_{K+1})$$

Similarly, define $F_{K+1} : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}^{r_1 \cdots r_K}$ as $F_{K+1}(\mathcal{A}, \mathcal{A}') = \nabla_{\mathcal{C}} \mathcal{L}(\mathcal{A})$. Let $F = (F_1, \dots, F_{K+1})$. Using F, define $G : \mathbb{R}^d \mapsto \mathbb{R}^d$ as:

$$G(\mathcal{A}) \triangleq F(S(\mathcal{A}), \mathcal{A}).$$

Intuitively, k-th block component of G can be considered as the partial derivative for M_k of $\mathcal{L}$, given the half-step iterate after updating M_k. Because each block update exists a unique solution, $G(\mathcal{A}) = 0$ holds in the neighborhood of $(\mathcal{A}^*)$. Differentiate the both side of $G(\mathcal{A}^*) = 0$, then we have:

$$\nabla G(\mathcal{A}^*) = \nabla_{\mathcal{A}} F(S(\mathcal{A}^*), \mathcal{A}) \nabla S(\mathcal{A}^*) + \nabla_{\mathcal{A}'} F(S(\mathcal{A}^*), \mathcal{A}^*) = 0 \tag{2}$$

To solve $\nabla S(\mathcal{A}^*)$, the Hessian of $\mathcal{L}$ at $\mathcal{A}^*$ is:

$$H(\mathcal{A}^*) = \nabla^2 \mathcal{L}(\mathcal{C}^*, M_1^*, \cdots, M_K^*) = \begin{pmatrix} d^2_{CC}\mathcal{L} & d^2_{CM_1}\mathcal{L} & \cdots & d^2_{CM_K}\mathcal{L} \\ d^2_{M_1C}\mathcal{L} & d^2_{M_1M_1}\mathcal{L} & \cdots & d^2_{M_1M_K}\mathcal{L} \\ \vdots & \vdots & \ddots & \vdots \\ d^2_{M_KC}\mathcal{L} & d^2_{M_KM_1}\mathcal{L} & \cdots & d^2_{M_KM_K}\mathcal{L} \end{pmatrix} = L + D + L^\top,$$

Since $H(\mathcal{A}^*)$ is strictly negative-definite except the direction of orthogonal transformation, the diagonal block of $H(\mathcal{A}^*)$, D, is strictly negative definite and thus $(L+D)^{-1}$ invertible. Reorganized the equation2, we can get $\nabla S(\mathcal{A}^*) = -(L+D)^{-1} L^T$.

Next, we construct the contraction relationship between iterates $\mathcal{B}^{(t+1)}$ and $\mathcal{B}^{(t)}$ using ∇S. For simplicity, let $\|\mathcal{A}, \mathcal{A}'\|_F$ denote the euclidean distance between two decision variables, where

$$\|\mathcal{A} - \mathcal{A}'\|_F = \|\mathcal{C} - \mathcal{C}'\|_F + \sum_{k=1}^K \|M_k - M'_K\|_F.$$

And we define the orthogonal transformation of $\mathcal{A}$. If $\mathcal{A}'$ is an orthogonal transformation of $\mathcal{A}$, there are orthogonal matrices $\{P_k\} \in \mathbb{O}_{r_k}$ such that:

$$M_k \cdot P_k^T = M_k^*, \forall k \in [K]; \quad \mathcal{C}^{(t)} \times_1 P_1 \times_2 \cdots \times_K P_K = \mathcal{C}^*; \quad \Rightarrow \mathcal{B}(\mathcal{A}) = \mathcal{B}(\mathcal{A}')$$

In our context, let $\mathcal{A} \in \Omega_O$ if $\mathcal{A}$ is an orthogonal transformation of $\mathcal{A}^*$, otherwise let $\mathcal{A} \in \Omega$. If $\mathcal{A} \in \Omega_O$, then $\mathcal{A} - \mathcal{A}^*$ is a direction that tangent to the orthogonal transformation of $\mathcal{A}^*$.

Here, we discuss two cases.

Case 1: The iterate $\mathcal{A}^{(t)} \in \Omega_O$.

For such $\mathcal{A}^{(t)}$, we have $\mathcal{B}(\mathcal{A}^{(t)}) = \mathcal{B}(\mathcal{A}^*)$. Trivially,

$$\left\|\mathcal{B}(\mathcal{A}^{(t)}) - \mathcal{B}(\mathcal{A}^*)\right\|_F = 0 \leq \left\|\mathcal{A}^{(0)} - \mathcal{A}^*\right\|_F, \tag{3}$$

for any $\mathcal{A}^{(0)}$.

Case 2: The iterate $\mathcal{A}^{(t)} \in \Omega$.

Therefore, $\mathcal{A}^{(t)} - \mathcal{A}^*$ is not on a direction that tangent to the orthogonal transformation of $\mathcal{A}^*$ and thus $H(\mathcal{A}^*)$ is strictly negative definite on the direction $\mathcal{A}^{(t)} - \mathcal{A}^*$. For $\forall \mathcal{A}^{(t)} \in \Omega$, we have:

$$(\mathcal{A}^{(t)} - \mathcal{A}^*)^T H(\mathcal{A}^*)(\mathcal{A}^{(t)} - \mathcal{A}^*) < 0 \tag{4}$$

Consider the matrix $\nabla S(\mathcal{A}^*)^T H(\mathcal{A}^*)\nabla S(\mathcal{A}^*) - H(\mathcal{A}^*)$. Let $H, \nabla S$ be the short of $H(\mathcal{A}^*), \nabla S(\mathcal{A}^*)$. We have:

$$\begin{aligned}
\nabla S(\mathcal{A}^*)^T H(\mathcal{A}^*)\nabla S(\mathcal{A}^*) - H(\mathcal{A}^*) &= \nabla S H \nabla S - H \\
&= (I - (L+D)^{-1}H)^T H (I - (L+D)^{-1}H) - H \\
&= -H^T (L+D)^{-1,T} H - H(L+D)^{-1}H + H^T(L+D)^{-1,T}H(L+D)^{-1}H \\
&= H^T(L+D)^{-1,T}\{-(L+D) - (L+D)^T + H\}(L+D)^{-1}H \\
&= H^T(L+D)^{-1,T}\{-D\}(L+D)^{-1}H
\end{aligned} \tag{5}$$

Since D is negative definite, then $-D$ is positive definite. For arbitrary $\mathcal{A}^{(t)} \in \Omega$, let $v \triangleq \mathcal{A}^{(t)} - \mathcal{A}^*$. Due to equation , $Hv \neq 0$. Multiplying v on both side of equation 5 , we have:

$$\begin{aligned}
v^T(\nabla S H \nabla S - H)v &= v^T H^T (L+D)^{-1,T}\{-D\}(L+D)^{-1}Hv > 0 \\
\Rightarrow \quad -v^T H v &> -(\nabla S v)^T H (\nabla S v)
\end{aligned}$$

Pick a v which is an eigenvector of ∇S with eigenvalue λ, then :

$$-v^T H v > -\lambda^2 v^T H v; \quad \Rightarrow \lambda^2 < 1$$

That implies, for $\mathcal{A}^{(t)} \in \Omega$, the largest eigenvalue of ∇S that corresponds to eigenvectors in form of $\mathcal{A}^{(t)} - \mathcal{A}^*$ is smaller than 1. Therefore, $\|\nabla S(\mathcal{A}^{(t)} - \mathcal{A}^*)\|_F \leq \rho \|\mathcal{A}^{(t)} - \mathcal{A}^*\|_F$ for $\forall \mathcal{A}^{(t)} \in \Omega$, where $\rho \in (0,1)$.

This is my conjecture. Intuitively, I want to show the spectral radius of S' < 1 in the space Omega\Omega_O. Then use the result that Fnorm(Ax)\leq \rho(A) Fnorm(x).

Consider the iterate $\mathcal{A}^{(t)} \in \Omega$, we have

$$\begin{aligned}
\left\|S(\mathcal{A}^{(t)}) - S(\mathcal{A}^*)\right\|_F &= \left\|\int_0^1 \nabla S(\mathcal{A}^* - u(\mathcal{A}^* - \mathcal{A}^{(t)}))(\mathcal{A}^* - \mathcal{A}^{(t)})du\right\|_F \\
&\leq \int_0^1 \left\|\nabla S(\mathcal{A}^* - u(\mathcal{A}^* - \mathcal{A}^{(t)}))(\mathcal{A}^* - \mathcal{A}^{(t)})\right\|_F du.
\end{aligned} \tag{6}$$

Since $\nabla S(\mathcal{A})$ is continuous and $\rho < 1$, pick a $\epsilon > 0$ such that $\epsilon + \rho < 1$, there exists a $\delta > 0$ such that

$$\text{If } \left\|\mathcal{A}^* - u(\mathcal{A}^* - \mathcal{A}^{(t)}) - \mathcal{A}^*\right\|_F \le \left\|\mathcal{A}^{(t)} - \mathcal{A}^*\right\|_F \le \delta, \text{ then } \left\|\nabla S - \nabla S(\mathcal{A}^* - u(\mathcal{A}^* - \mathcal{A}^{(t)})\right\|_F \le \epsilon$$

Therefore, the inequality 6 becomes:

$$\begin{aligned}\left\|S(\mathcal{A}^{(t)}) - S(\mathcal{A}^*)\right\|_F &\le \int_0^1 \left\|\nabla S(\mathcal{A}^* - u(\mathcal{A}^* - \mathcal{A}^{(t)}))(\mathcal{A}^* - \mathcal{A}^{(t)})\right\|_F du.\\ &\le (\rho + \epsilon)\left\|\mathcal{A}^{(t)} - \mathcal{A}^*\right\|_F\end{aligned}$$

If any previous iterate $\mathcal{A}^{(t')}, t' < t$ is not in Ω_O, then we have :

$$\left\|\mathcal{A}^{(t)} - \mathcal{A}^*\right\|_F \le \rho^t \left\|\mathcal{A}^* - \mathcal{A}^{(0)}\right\|_F,$$

for $\mathcal{A}^{(0)}$ sufficiently closes to $\mathcal{A}^*$ and is not a local maximizer. By the Lemma 3.1 of Han[2020], there exists a constant c such that:

$$\left\|\mathcal{B}(\mathcal{A}^{(t)}) - \mathcal{B}(\mathcal{A}^*)\right\|_F \le c\left\|\mathcal{A}^{(t)} - \mathcal{A}^*\right\|_F. \tag{7}$$

If there exists a iterate $\mathcal{A}^{(t')}, t' < t$ such that $\mathcal{A}^{(t')} \in \Omega_O$, then we goes to case 1.

Combine the equation 3 and 7 , we can summarize our local convergence as:

$$\left\|\mathcal{B}(\mathcal{A}^{(t)}) - \mathcal{B}(\mathcal{A}^*)\right\|_F \le c\rho^t \left\|\mathcal{A}^* - \mathcal{A}^{(0)}\right\|_F,$$

for some constant c and $\mathcal{A}^{(0)}$ sufficiently closes to $\mathcal{A}^*$ and is not a local maximizer.