TR Convergence Proof Difference

Jiaxin Hu

06/12/2020

1 LOCAL CONVERGENCE

Proposition 1 (Local Convergence)**.** *Assume the solution to each block update in the alternating optimization exists and is unique. Let $\mathcal{B}^* = (\mathcal{C}^*, \{\boldsymbol{M}_1^*, \dots, \boldsymbol{M}_K^*\})$ be a local minimizer of $\mathcal{L}_{\mathcal{Y}}$ and assume the Hessian ~~at $\mathcal{B}^*$~~ is strictly negative definite ~~in every direction except those tangent to~~ with respect to the block variables module the orthogonal transformation of $\boldsymbol{M}_k^*$. Then the sequence $\mathcal{B}^{(t)} = \mathcal{C}^{(t)} \times \{\boldsymbol{M}_1^*, \dots, \boldsymbol{M}_K^*\}$ generated by alternating algorithm linearly converges to $\mathcal{B}^*$; i.e.*

$$\left\|\mathcal{B}^{(t)} - \mathcal{B}^*\right\|_F \le \rho^t \Big(\left\|\mathcal{C}^{(0)} - \mathcal{C}\right\|_F + \sum_{k=1}^{K} \left\|\boldsymbol{M}_k^{(0)} - \boldsymbol{M}_K^*\right\|_F\Big),$$

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

PROOF

For notational convenience, we drop the subscript $\mathcal{Y}$ from the objective $\mathcal{L}_{\mathcal{Y}}(\cdot)$ and simply write as $\mathcal{L}(\cdot)$. Let $\mathcal{A} = (\mathcal{C}, \boldsymbol{M}_1, \dots, \boldsymbol{M}_K) \in \mathbb{R}^{d_{\text{total}}}$ denote the collection of decision variables used in the alternating optimization, where $d_{\text{total}} = \prod_k r_k + \sum_k r_k d_k$. The objective function can be viewed either as a function of decision variables $\mathcal{A}$ or a function of coefficient tensor $\mathcal{B} := \mathcal{C} \times_1 \boldsymbol{M}_1 \times_2 \cdots \times_K \boldsymbol{M}_K$. With slight abuse of notation, we write both functions as $\mathcal{L}(\cdot)$ but the meaning should be clear given the context.

~~Let~~ We use $S : \mathbb{R}^{d_{total}} \mapsto \mathbb{R}^{d_{total}}$ denote the update mapping that sends the 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, we have $S(\mathcal{A}^{(t)}) = \mathcal{A}^{(t+1)}$ ~~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~~ a composition of $K+1$ block-wise mappings. Each block-wise mapping is continuously differentiable, so the mapping S is also continuously differentiable. ~~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 $\{M_k\}$:~~

$$C_k(\mathcal{C}, M_1, \dots, M_{k-1}, M_{k+1}, \dots, M_K) \overset{\Delta}{=} C_k, \text{ where } \nabla_{M_k} \mathcal{L}(\mathcal{C}, M_1, \dots, M_{k-1}, C_k, M_{k+1}, \dots, M_K) = 0 \tag{1}$$

$$\text{and } C_{K+1}(\{M_k\}) \overset{\Delta}{=} C_{K+1}, \text{ where } \nabla_{\mathcal{C}} \mathcal{L}(C_{K+1}, \{M_k\}) = 0.$$

~~Because each block update exists a unique solution, there exists such a C_k satisfies the condition 1 and $\nabla_{M_k,M_k}\mathcal{L}(\mathcal{C}, M_1, \ldots, M_{k-1}, C_k, M_{k+1}, \ldots, M_K)$ 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:~~

$$S_k(\mathcal{C}, \{M_k\}) \triangleq (\mathcal{C}, M_1, \ldots, M_{k-1}, C_k, M_{k+1}, \ldots, M_K), \forall k \in [K]$$

$$S_{K+1}(\mathcal{C}, \{M_k\}) \triangleq (C_{K+1}, \{M_k\})$$

~~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:~~

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

~~Similarly, define $F_{K+1} : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}^{r_1 \ldots r_K}$ as $F_{K+1}(\mathcal{A}, \mathcal{A}') = \nabla_{\mathcal{C}}\mathcal{L}(\mathcal{A})$. Let $F = (F_1, \ldots, F_{K+1})$. 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:~~ Let $\mathcal{A}^* = (\mathcal{C}^*, \boldsymbol{M}_1^*, \ldots, \boldsymbol{M}_K^*) \in \mathbb{R}^{d_{\text{total}}}$ be a local maximum. By the definition of alternating optimization, $\mathcal{A}^*$ is also a fixed point for the mapping S; that is, $S(\mathcal{A}^*) = \mathcal{A}^*$. The Hessian of the objective function $\mathcal{L}(\cdot)$ at $\mathcal{A}^*$ is

$$H(\mathcal{A}^*) = \nabla^2\mathcal{L}\left(\mathcal{C}^*, \boldsymbol{M}_1^*, \cdots, \boldsymbol{M}_K^*\right) = \begin{pmatrix} \nabla^2_{CC}\mathcal{L} & \nabla^2_{CM_1}\mathcal{L} & \cdots & \nabla^2_{CM_K}\mathcal{L} \\ \nabla^2_{M_1C}\mathcal{L} & \nabla^2_{M_1M_1}\mathcal{L} & \cdots & \nabla^2_{M_1M_K}\mathcal{L} \\ \vdots & \vdots & \ddots & \vdots \\ \nabla^2_{M_KC}\mathcal{L} & \nabla^2_{M_KM_1}\mathcal{L} & \cdots & \nabla^2_{M_KM_K}\mathcal{L} \end{pmatrix} =: L + D + L^\top,$$

where D collects the diagonal blocks and L collects the lower-diagonal blocks. By assumption ~~Since~~, $H(\mathcal{A}^*)$ is strictly negative definite at every direction except the direction of orthogonal transformation. That implies that the digaonal block ~~, the diagonal block of $H(\mathcal{A}^*)$,~~ D, is strictly negative definite and thus $(L+D)^{-1}$ invertible. By [Bezdek, 2003, Lemma2], we have ~~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 the property of ∇S in the neighborhood of $\mathcal{A}^*$.

We need to introduce some additional notation. ~~For simplicity,~~ Let $\|\mathcal{A}-\mathcal{A}'\|_F$ denote the E~~e~~uclidean distance between two decision variables, where

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

~~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:~~ We introduce the equivalent relationship induced by orthogonal transformation. Let $\mathbb{O}_{d,r}$ be the collection of all d-by-r matrices with orthogonal columns, $\mathbb{O}_{d,r} := \{\boldsymbol{P} \in \mathbb{R}^{d\times r} : \boldsymbol{P}^T\boldsymbol{P} = \mathbf{1}_r\}$, where $\mathbf{1}_r$ is the r-by-r identity matrix.

Definition 1 (Equivalence relationship)**.** *Two decision variables $\mathcal{A}' = (\mathcal{C}', \boldsymbol{M}'_1, \ldots, \boldsymbol{M}'_k)$, $\mathcal{A} = (\mathcal{C}, \boldsymbol{M}_1, \ldots, \boldsymbol{M}_k)$ are called equivalent, denoted $\mathcal{A} \sim \mathcal{A}'$, if and only if there exist a set of orthogonal matrices $\boldsymbol{P}_k \in \mathbb{O}_{d_k, r_k}$ such that*

$$\boldsymbol{M}_k^{(t)}\boldsymbol{P}_k^T = \boldsymbol{M}_k^*, \forall k \in [K]; \quad \mathcal{C}^{(t)} \times_1 \boldsymbol{P}_1 \times_2 \cdots \times_K \boldsymbol{P}_K = \mathcal{C}^*; \quad \Rightarrow \mathcal{B}(\mathcal{A}) = \mathcal{B}(\mathcal{A}')$$

~~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 $\mathcal{A}^*$.~~ Equivalently, two decision variables $\mathcal{A}$, $\mathcal{A}'$ are equivalent if the corresponding Tucker tensors are the same, $\mathcal{B}(\mathcal{A}) = \mathcal{B}'(\mathcal{A}')$. We use Ω_O to denote all decision variables that are equivalent to the local optimum $\mathcal{A}^*$, $\Omega_O := \{\mathcal{A} \in \mathbb{R}^{d_{\text{total}}} : \mathcal{A} \sim \mathcal{A}^*\}$. Here, we discuss two cases at a sufficiently-small neighborhood of $\mathcal{A}^*$. ~~Here, we discuss two cases.~~

Case 1: ~~The iterate $\mathcal{A}^{(t)} \in \Omega_O$.~~ There exists an iteration number $t' \in \mathbb{N}_+$ such that $\mathcal{A}^{(t')} \in \Omega_O$. For such $\mathcal{A}^{(t')}$, we have $\mathcal{B}(\mathcal{A}^{(t')}) = \mathcal{B}(\mathcal{A}^*)$. Therefore, ~~Trivially,~~

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

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

Case 2: ~~The iterate $\mathcal{A}^{(t)} \in \Omega$.~~ The entire sequence of iterates $\mathcal{A}^{(t)} \in \mathbb{R}^{d_{total}}/\Omega_O$. By assumption, $H(\cdot)$ is strictly negative definite for all t large enough. ~~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 any such $\forall \mathcal{A}^{(t)} \in \Omega$, 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 1, $Hv \neq 0$. Multiplying v on both side of equation 5 , we have:~~

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

~~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 $\rho \in (0,1)$.~~ Recall that the differential map $\nabla S(\mathcal{A}^*) = -(L+D)^{-1} L^T$, where L, D are the lower- and diagonal-block of the Hession $H(\mathcal{A}^*)$, respectively. Define contraction coefficient

$$
\rho = \max_{\boldsymbol{x} \in \mathbb{R}^{d_{\text{total}}}/\Omega_O,\ \|\boldsymbol{x}\|_2 = 1} \boldsymbol{x}^T \left[(L+D)^{-1} L \right] \in (0,1).
$$

~~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~~

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

~~Therefore, the inequality 6 becomes:~~

$$\cancel{\left\| S(\mathcal{A}^{(t)}) - S(\mathcal{A}^*) \right\|_F \leq \int_0^1 \left\| \nabla S(\mathcal{A}^* - u(\mathcal{A}^* - \mathcal{A}^{(t)}))(\mathcal{A}^* - \mathcal{A}^{(t)}) \right\|_F du.}$$
$$\cancel{\leq (\rho + \epsilon) \left\| \mathcal{A}^{(t)} - \mathcal{A}^* \right\|_F}$$

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

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

By the contraction principle, we have

$$\|\mathcal{A}^{(t)} - \mathcal{A}^*\|_F \leq \rho^t \|\mathcal{A}^{(0)} - \mathcal{A}^*\|_F, \tag{7}$$

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

$$\left\| \mathcal{B}(\mathcal{A}^{(t)}) - \mathcal{B}(\mathcal{A}^*) \right\|_F \leq c \left\| \mathcal{A}^{(t)} - \mathcal{A}^* \right\|_F, \quad \forall t \in \mathbb{N}_+. \tag{8}$$

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

Combining ~~Combine the equation~~ 3 and 8 , ~~we can summarize our local convergence as:~~

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

for ~~some constant c and~~ $\mathcal{A}^{(0)}$ sufficiently closes to $\mathcal{A}^*$ ~~and is not a local maximizer~~. Combining cases 1 and 2, we obtain that

$$\|\mathcal{B}(\mathcal{A}^{(t)}) - \mathcal{B}(\mathcal{A}^*)\|_F^2 \leq c\rho^{2t} \left(\|\mathcal{C}^{(0)} - \mathcal{C}^*\|_F^2 + \sum_{k=1}^{K} \|\boldsymbol{M}_k^{(0)} - \boldsymbol{M}_k^*\|_F^2 \right),$$

for some constant $c > 0$ and any initialization $\mathcal{A}^{(0)} = (\mathcal{C}^{(0)}, \boldsymbol{M}_1^{(0)}, \ldots, \boldsymbol{M}_K^{(0)})$ sufficiently close to $\mathcal{A}^* = (\mathcal{C}^*, \boldsymbol{M}_1^*, \ldots, \boldsymbol{M}_K^*)$.

2 GLOBAL CONVERGENCE

Proposition 2 (Global Convergence)**.** *Assume the set $\{\mathcal{A} : \mathcal{L}(\mathcal{A}) \geq \mathcal{L}(\mathcal{A}^{(0)})\}$ is compact and the stationary points of $\mathcal{L}(\mathcal{A})$ are isolated module equivalence. Then any sequence $\mathcal{A}^{(t)}$ generated by alternating algorithm converges to a stationary point of $\mathcal{L}(\mathcal{A})$ up to equivalence.*

PROOF

Pick an arbitrary iterate $\mathcal{A}^{(t)}$. Because of the compactness of set $\{\mathcal{A}: \ \mathcal{L}(\mathcal{A}) \geq \mathcal{L}(\mathcal{A}^{(0)})\}$ and the boundedness of the decision domain , ~~the domain of $\mathcal{A}^{(t)}$ is bounded and thus~~ there exists~~s convergent~~ a sub-sequence~~s~~ of $\mathcal{A}^{(t)}$ that converges. Let $\mathcal{A}^*$ denote ~~a~~ one of the limiting points of $\mathcal{A}^{(t)}$. ~~Since $\mathcal{L}(\mathcal{A}^{(t)})$~~

~~increases monotonically along with $t \to \infty$, then $\mathcal{A}^*$ is a stationary point of $\mathcal{L}(\mathcal{A})$.~~ Let $\mathcal{S} = \{\mathcal{A}^*\}$ denote the set of all the limiting points of $\mathcal{A}^{(t)}$. We have $\mathcal{S} \subset \{\mathcal{A} | \; \mathcal{L}(\mathcal{A}) \geq \mathcal{L}(\mathcal{A}^{(0)})\}$ and thus $\mathcal{S}$ is a compact set. ~~According to~~ By [Lange, 2012], $\mathcal{S}$ is also connected. Note that all points in $\mathcal{S}$ are also stationary points of $\mathcal{L}(\cdot)$, because of the monotonic increase of $\mathcal{L}(\mathcal{A}^{(t)})$ as $t \to \infty$.

Consider the equivalence of Tucker tensor representation. We define the equivalent class of $\mathcal{A}$ as:

$$\mathcal{E}(\mathcal{A}) = \{\mathcal{A}' | \; \boldsymbol{M}'_k = \boldsymbol{M}_k \boldsymbol{P}_k^T, \mathcal{C}' = \mathcal{C} \times \{\boldsymbol{P}_1, \ldots, \boldsymbol{P}_K\}, \text{ where } \boldsymbol{P}_k^T \in \mathbb{O}_{d_k, r_k}, \forall k \in [K]\}.$$

~~Notice that, for arbitrary $\mathcal{A}$, $\mathcal{E}(\mathcal{A})$ is a non-empty open set. For arbitrary two non-equivalent points $\mathcal{A}_1$ and $\mathcal{A}_2$, we have $\mathcal{E}(\mathcal{A}_1) \cap \mathcal{E}(\mathcal{A}_2) = \varnothing$ and thus $\mathcal{E}(\mathcal{A}_1) \cup \mathcal{E}(\mathcal{A}_2)$ is not connected. Using the definition of equivalent class, let $\mathcal{S}_E$ denote the enlarged set of $\mathcal{S}$, such that:~~

We define an enlarges set $\mathcal{E}_S$ induced by the set $\mathcal{S}$,

$$\mathcal{S}_E \mathcal{E}_S = \bigcup_{\mathcal{A} \in \mathcal{S}} \{\mathcal{E}(\mathcal{A}^*) : \mathcal{A}^* \in \mathcal{S}\}.$$

The enlarged set $\mathcal{S}_E$ satisfies ~~below~~ two properties below :

1. [Union of Stationary Point] The set ~~$\mathcal{S}_E$~~ $\mathcal{E}_S$ is an union of equivalent classes generated by the stationary points in $\mathcal{S}$.

2. [Connectedness model the equivalence] The set ~~$\mathcal{S}_E$~~ $\mathcal{E}_S$ is connected ~~between different equivalent classes~~ module the equivalence relationship. That property is obtained by the connectedness of $\mathcal{S}$

~~Property 1 is obtained by rewriting the definition of $\mathcal{S}_E$. Property 2 is concluded by the connectedness of $\mathcal{S}$.~~

Now, note that the isolation of stationary points and Property 1 imply that ~~$\mathcal{S}_E$~~ $\mathcal{E}_S$ ~~only~~ contains only finite number of ~~different~~ equivalent classes. Otherwise, there is a sequence of non-equivalent stationary points whose limit is not isolated, which contradicts the isolation assumption. ~~Combined~~ Combining the finiteness with ~~the definition of equivalent class and~~ Property 2, we ~~can~~ conclude that ~~$\mathcal{S}_E$~~ $\mathcal{E}_S$ ~~only~~ contains only a single equivalent class; i.e. $\mathcal{E}_S$ $\mathcal{S}_E$ $= \mathcal{E}(\mathcal{A}^*)$, where $\mathcal{A}^*$ is a stationary point of $\mathcal{L}(\mathcal{A})$. Therefore, all the convergent sub-sequences of $\mathcal{A}^{(t)}$ converge to one stationary point $\mathcal{A}^*$ up to equivalence. ~~In other words,~~ We conclude that, any iterate $\mathcal{A}^{(t)}$ generated by Algorithm 1 converges to a stationary point of $\mathcal{L}(\mathcal{A})$ up to equivalence.