Convergence Property for Tensor Regression

Jiaxin Hu

05/24/2020

CONVERGENCE PROPERTY

The convergence of Algorithm 1 is guaranteed because the alternating algorithm monotonically increase the objective function after each iteration. Now, we study the convergence property for the actual estimation sequence $(\mathcal{C}^{(t)}, \{M_k^{(t)}\})$ and $\mathcal{B}^{(t)} = \mathcal{C}^{(t)} \times_1 M_1^{(t)} \times_2 \cdots \times_K M_K^{(t)}$. To simplify the analysis, we set the hyper-parameter α to infinity and define the notation $\mathcal{A} = (\mathcal{C}, \{M_k\})$. Define the Forbenius norm of $\|\mathcal{A}\|_F = \|\mathcal{C}\|_F + \sum_{k=1}^K \|M_k\|_F$. We propose following assumptions.

Is it true that the Frobenius norm of M_k is always \sqrt{r_k}?

A1.(Regularity Condition) The log-likelihood function $\mathcal{L}(\mathcal{A})$ is continuous and the set $\{\mathcal{A} : \mathcal{L}(\mathcal{A}) \geq \mathcal{L}(\mathcal{A}^{(0)})\}$ is compact.

The Hessian of what function w.r.t. what decision variables?

A2.(Strictly local maximum condition) Each block update in Algorithm is well-defined and the corresponding Hessian is non-singular at the solution.

A3.(Local Uniqueness condition) The set of stationary points of $\mathcal{L}(\mathcal{A})$ is isolated up to orthogonalization.

A4.(Tensor Lipschitz condition) The tensor representation $\mathcal{B}(\mathcal{A})$ is tensor Lipschitz at $\mathcal{A}^*$. That means there exists two constant $c_1, c_2 > 0, s.t.$

Does this assumption hold in our context?

$$c_1 \left\|\mathcal{A}' - \mathcal{A}''\right\|_F \leq \left\|\mathcal{B}(\mathcal{A}') - \mathcal{B}(\mathcal{A}'')\right\|_F \leq c_2 \left\|\mathcal{A}' - \mathcal{A}''\right\|_F$$

What if A', A'' differ by orthogonalization? B(A')-B(A'') = 0, but A' \neq A''. Consider the following couter-example
A'=[C; M1,…MK];
A''=[C\times_1 P; M1*P^{-1}, M2,…,MK]

for $\mathcal{A}', \mathcal{A}''$ are sufficiently close to $\mathcal{A}^*$.

These conditions are mild for the tensor of order 3 or higher.

Proposition 1 (Algorithm 1 Convergence)**.** *Suppose A1-A3 holds.*

1. (Global Convergence) Any sequence $\mathcal{A}^{(t)}$ generated by Algorithm 1 convergences to a stationary point of $\mathcal{L}(\mathcal{A})$.

2.(Local Linear Convergence) Let $\mathcal{A}^*$ be a local maximizer of $\mathcal{L}(\mathcal{A})$. There exists an ϵ-neighborhood of $\mathcal{A}^*$, such that, for any $\mathcal{A}^{(0)}$ in the neighborhood, the iterates $\mathcal{A}^{(t)}$ generated by Algorithm 1 linearly convergent to $\mathcal{A}^*$.

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

where $\rho \in (0,1)$ is a contraction parameter. If A4 holds, there exists a constant C such that

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

Combining the Proposition and Theorem 4.1, we have the empirical performance of our estimate.

Theorem 1 (Empirical Performance). *Let $\mathcal{A}^{(t)}$ be a sequence of estimates generated by Algorithm 1 with initial point $\mathcal{A}^{(0)}$ and limiting point $\mathcal{A}^*$. Suppose $Loss(\mathcal{B}^{\mathcal{A}^{(0)}}, \mathcal{A}_{true}) = \epsilon$ and $\mathcal{L}(\mathcal{A}^*) > \mathcal{L}(\mathcal{A}_{true})$. Suppose A1-A4 holds. With probability $1 - exp(-C_1 \sum_k p_k)$,*

typo

$$Loss(\mathcal{B}^{\mathcal{A}^{(t)}}, \mathcal{A}_{true}) \le C\rho^t \epsilon + C_2 \sum_k p_k,$$

where $\rho \in (0,1)$ is a contraction parameter and $C_1, C_2 > 0$ are two constants.

PROOF

gramma error

For Global Convergence, we need to show every sub-sequence of $\mathcal{A}^{(t)}$ accumulates to the same stationary point. Let $\mathcal{A}^*$ be any limiting point of sub-sequences of $\mathcal{A}^{(t)}$. Due to $\mathcal{L}(\mathcal{A}^{(t)})$ monotonically increases along with t, every $\mathcal{A}^*$ is a stationary point of $\mathcal{L}(\mathcal{A})$ and $\mathcal{A}^* \in \{\mathcal{A} : \mathcal{L}(\mathcal{A}) \ge \mathcal{L}(\mathcal{A}^{(0)})\}$. The set of $\mathcal{A}^*$ is also compact and connected because of *A1*. The isolation of stationary point also implies there are only finite $\mathcal{A}^*$s. Therefore, the set of $\mathcal{A}^*$ becomes a single point. The Global Convergence holds.

To show the Local Convergence, define the differential mapping $S : S(\mathcal{A}^{(t)}) = \mathcal{A}^{(t+1)}$. Let H be the Hessian matrix of $\mathcal{L}(\mathcal{A})$ at the local maximum $\mathcal{A}^*$. We partition the H:

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

where L is strictly block lower triangle matrix and D is the diagonal part. Due to *A2*, every sub-matrix of Hessian is non-singular and negative definite. Therefore, $L + D$ is invertible. According to Bezdek(2003), we have $dS(\mathcal{A}^*) = -(L+d)^{-1}L$. Let ρ denote the spectral radius of $dS(\mathcal{A}^*)$ and $\rho = max_i|\lambda_i(-(L+d)^{-1}L)|$ is strictly smaller than 1. According to contraction principle, Algorithm 1 will be linearly convergent if its spectral radius of $dS(\mathcal{A}^*)$ is smaller than 1. Therefore,

$$d(S(\mathcal{A}^{(t)}), S(\mathcal{A}^*)) \le \rho^t d(\mathcal{A}^{(0)}, \mathcal{A}^*), \quad d(S(\mathcal{A}), S(\mathcal{A}')) = \left\|\mathcal{C} - \mathcal{C}'\right\|_F + \sum_k \left\|M_k - M_k'\right\|_F$$

where $(\mathcal{A}, d)$ is a complete metric space, $S(\mathcal{A}^*) = \mathcal{A}^*$ and $\mathcal{A}^{(0)}$ is sufficiently near to $\mathcal{A}^*$. With sufficiently large $t \in \mathbb{N}^+$ and *A4*, we have

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

where $C > 0$ is a constant.