Modern Hopfield Network
Global Convergence Theory Explained

Jiongjiong Li
jiongjiongfly@gmail.com

October 14, 2024

Contents

1 Introduction

This article provides a comprehensive explanation of the global convergence theory of the Modern Hopfield Network [1], covering energy and stationary points. It also extends the discussion of the initial point $\boldsymbol{\xi}_0$ from a bounded subset to any vector in $\mathbb{R}^d$.

2 Notations

1. The p-norm of a vector is denoted: $\|\cdot\|_p$,$1 \leqslant p$
2. Let $\beta > 0, \beta \in \mathbb{R}$
3. Let $[N] = \{1, \cdots, N\}$

3 Softmax Function

$$\mathbf{p}(\mathbf{x}) = \mathrm{softmax}(\beta \mathbf{x})$$
$$p_i = [\mathrm{softmax}(\beta \mathbf{x})]_i = \frac{\exp(\beta x_i)}{\sum_{k=1}^{N} \exp(\beta x_k)} \tag{1}$$

where $\mathbf{x}_i \in \mathbb{R}^N$.

4 Log-Sum-Exp Function

$$\mathrm{lse}(\beta, \mathbf{x}) = \beta^{-1} \ln \left(\sum_{k=1}^{N} \exp(\beta x_k) \right) \tag{2}$$

where $\mathbf{x}_i \in \mathbb{R}^N$.

5 Lemma 1 LSE Gradient

$$\nabla_{\mathbf{x}} \mathrm{lse}(\beta, \mathbf{x}) = \mathrm{softmax}(\beta \mathbf{x}) \tag{3}$$

where $\mathbf{x} \in \mathbb{R}^N$.

$$\nabla_{\boldsymbol{\xi}} \mathrm{lse}\left(\beta, \boldsymbol{X}^T \boldsymbol{\xi}\right) = \boldsymbol{X} \,\mathrm{softmax}\left(\beta \boldsymbol{X}^T \boldsymbol{\xi}\right) \tag{4}$$

where $\boldsymbol{X} \in \mathbb{R}^{d \times N}$, $\boldsymbol{\xi} \in \mathbb{R}^d$.

6 Lemma 2 LSE Hessian

The Jacobian $J_s(\mathbf{x})$ of $\mathbf{p}(\mathbf{x}) = \mathrm{softmax}(\beta \mathbf{x})$ is

$$J_s(\mathbf{x}) = \frac{\partial \,\mathrm{softmax}(\beta \mathbf{x})}{\partial \mathbf{x}} = \beta \left(\mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^T \right) \tag{5}$$

where $\mathbf{x} \in \mathbb{R}^N$.
The Jacobian $J(\boldsymbol{\xi})$ of $\boldsymbol{X}\mathbf{p}\left(\boldsymbol{X}^T \boldsymbol{\xi}\right) = \boldsymbol{X} \,\mathrm{softmax}\left(\beta \boldsymbol{X}^T \boldsymbol{\xi}\right)$ is

$$J(\boldsymbol{\xi}) = \frac{\partial \left(\boldsymbol{X} \,\mathrm{softmax}\left(\beta \boldsymbol{X}^T \boldsymbol{\xi}\right)\right)}{\partial \boldsymbol{\xi}} = \beta \boldsymbol{X} \left(\mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^T \right) \boldsymbol{X}^T = \boldsymbol{X} J_s \boldsymbol{X}^T \tag{6}$$

where $\boldsymbol{X} \in \mathbb{R}^{d \times N}$, $\boldsymbol{\xi} \in \mathbb{R}^d$.

7 Lemma 3 LSE Hessian is Symmetric and Positive Semi-Definite

The Jacobian $J_s(\mathbf{x})$ of $\mathbf{p}(\mathbf{x}) = \mathrm{softmax}(\beta\mathbf{x})$ is symmetric and positive semi-definite where $\mathbf{x} \in \mathbb{R}^N$.
The Jacobian $J(\boldsymbol{\xi})$ of $\boldsymbol{X}\mathbf{p}\left(\boldsymbol{X}^T\boldsymbol{\xi}\right) = \boldsymbol{X}\,\mathrm{softmax}\left(\beta\boldsymbol{X}^T\boldsymbol{\xi}\right)$ is symmetric and positive semi-definite where $\boldsymbol{X} \in \mathbb{R}^{d\times N}, \boldsymbol{\xi} \in \mathbb{R}^d$.

Proof. According to 5, 6, they are symmetric. Given any $\mathbf{x} \in \mathbb{R}^N$, we have

$$\mathbf{x}^T J_s \mathbf{x} = \beta\mathbf{x}^T\left(\mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^T\right)\mathbf{x} = \beta\left[\sum_i p_i {x_i}^2 - \left(\sum_i p_i x_i\right)^2\right] \geqslant 0 \quad (7)$$

Given any $\boldsymbol{\xi} \in \mathbb{R}^d$, we have

$$\boldsymbol{\xi}^T J\boldsymbol{\xi} = \boldsymbol{\xi}^T\boldsymbol{X}J_s\boldsymbol{X}^T\boldsymbol{\xi} = \left(\boldsymbol{X}^T\boldsymbol{\xi}\right)^T J_s\left(\boldsymbol{X}^T\boldsymbol{\xi}\right) \geqslant 0 \quad (8)$$

Hence they are positive semi-definite.

□

8 Lemma 4 LSE is Convex

$\mathrm{lse}(\beta, \mathbf{x})$ is convex with respect to $\mathbf{x}$ where $\mathbf{x} \in \mathbb{R}^N$.
$\mathrm{lse}\left(\beta, \boldsymbol{X}^T\boldsymbol{\xi}\right)$ is convex with respect to $\boldsymbol{\xi}$ where $\boldsymbol{X} \in \mathbb{R}^{d\times N}, \boldsymbol{\xi} \in \mathbb{R}^d$.

Proof. The Jacobian of the $\mathrm{softmax}$ is Hessian of the lse according to 3, 4. The Jacobian of the $\mathrm{softmax}$ is positive semi-definite according to Lemma 3. Therefore lse is convex.

□

9 Definitions

9.1 Point-to-set Map

A point-to-set map $\mathcal{A}$ from a set X into a set Y is defined as

$$\mathcal{A} : X \to \mathscr{P}(Y) \tag{9}$$

which assigns a subset of Y to each point of X, where $\mathscr{P}(Y)$ denotes the power set of Y.

9.2 Closed

Suppose X and Y are two topological spaces. A point-to-set map $\mathcal{A} : X \to \mathscr{P}(Y)$ is said to be closed at $\mathbf{x}_0 \in X$ if:

$$\left.\begin{array}{l} \mathbf{x}_t \to \mathbf{x}_0 \text{ as } t \to \infty \\ \mathbf{y}_t \to \mathbf{y}_0 \text{ as } t \to \infty \\ \mathbf{x}_t \in X \\ \mathbf{y}_t \in \mathcal{A}(\mathbf{x}_t) \end{array}\right\} \implies \mathbf{y}_0 \in \mathcal{A}(\mathbf{x}_0) \tag{10}$$

A point-to-set map $\mathcal{A}$ is said to be closed on $S \subset X$ if it is closed at every point of S.

9.3 Fixed Point

A fixed point of a point-to-set map

$$\mathcal{A} : X \to \mathscr{P}(X) \tag{11}$$

is any point $\mathbf{x} \in X$ for which

$$\mathcal{A}(\mathbf{x}) = \{\mathbf{x}\} \tag{12}$$

A generalized fixed point of $\mathcal{A}$ is any point $\mathbf{x} \in X$ for which

$$\mathbf{x} \in \mathcal{A}(\mathbf{x}) \tag{13}$$

9.4 Uniformly Compact

A point-to-set map $\mathcal{A} : X \to \mathscr{P}(Y)$ is said to be uniformly compact on X if there exists a compact set $H \subset Y$ independent of $\mathbf{x}$ such that:

$$\mathcal{A}(\mathbf{x}) \subset H, \forall \mathbf{x} \in X \tag{14}$$

9.5 Monotonic

Let

$$\mathrm{E} : X \to \mathbb{R} \tag{15}$$

be a continuous function. The map

$$\mathcal{A} : X \to \mathscr{P}(X) \tag{16}$$

is said to be monotonic with respect to E at $\mathbf{x} \in X$ if

$$\forall \mathbf{y} \in \mathcal{A}(\mathbf{x}), \mathrm{E}(\mathbf{y}) \leqslant \mathrm{E}(\mathbf{x}) \tag{17}$$

$\mathcal{A}$ is said to be strictly monotonic with respect to E at $\mathbf{x} \in X$ if

$$\mathcal{A}(\mathbf{x}) \neq \{\mathbf{x}\} \implies \forall \mathbf{y} \in \mathcal{A}(\mathbf{x}), \mathrm{E}(\mathbf{y}) < \mathrm{E}(\mathbf{x}) \tag{18}$$

9.6 Global Convergence

Let X be a set and $\boldsymbol{\xi}_0 \in X$ a given point. Then an algorithm, $\mathcal{A}$, with initial point $\boldsymbol{\xi}_0$ is a point-to-set map $\mathcal{A} : X \to \mathscr{P}(X)$ which generates a sequence $\{\boldsymbol{\xi}_t\}_{t=0}^{\infty}$ via the rule

$$\boldsymbol{\xi}_{t+1} \in \mathcal{A}(\boldsymbol{\xi}_t), t = 0, 1, \cdots \tag{19}$$

$\mathcal{A}$ is said to be globally convergent if: Given any chosen initial point $\boldsymbol{\xi}_0$, the sequence $\{\boldsymbol{\xi}_t\}_{t=0}^{\infty}$ generated by $\boldsymbol{\xi}_{t+1} \in \mathcal{A}(\boldsymbol{\xi}_t), t = 0, 1, \cdots$ (or a subsequence) converges to a point for which a necessary condition of optimality holds.

10 Modern Hopfield Network

We have patterns that are represented by the matrix:

$$\boldsymbol{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N) \tag{20}$$

where $\mathbf{x}_i \in \mathbb{R}^d$. Thus $\boldsymbol{X} \in \mathbb{R}^{d \times N}$. The largest norm of a pattern is:

$$M = \max_{i \in [N]} \|\mathbf{x}_i\|_2 \tag{21}$$

The query or state of the Hopfield Network is $\boldsymbol{\xi} \in \mathbb{R}^d$. Define energy $\mathrm{E}(\boldsymbol{\xi})$ for a continuous query or state $\boldsymbol{\xi}$:

$$\mathrm{E}(\boldsymbol{\xi}) = -\operatorname{lse}\left(\beta, \boldsymbol{X}^T\boldsymbol{\xi}\right) + \frac{1}{2}\boldsymbol{\xi}^T\boldsymbol{\xi} + \beta^{-1}\ln N + \frac{1}{2}M^2 \tag{22}$$

11 New Update Rule

$$\boldsymbol{\xi}_{t+1} = \mathcal{F}(\boldsymbol{\xi}_t) = \boldsymbol{X}\mathbf{p} = \boldsymbol{X}\operatorname{softmax}\left(\beta\boldsymbol{X}^T\boldsymbol{\xi}_t\right) \tag{23}$$

where

$$\mathbf{p} = \operatorname{softmax}\left(\beta\boldsymbol{X}^T\boldsymbol{\xi}_t\right) \tag{24}$$

12 Lemma 5 CCCP Algorithm

Consider an energy function $\mathrm{E} : X \to \mathbb{R}$:

$$\mathrm{E}(\boldsymbol{\xi}) = \mathrm{E}_1(\boldsymbol{\xi}) - \mathrm{E}_2(\boldsymbol{\xi}) \tag{25}$$

where E_1 and E_2 are differentiable convex functions of $\boldsymbol{\xi}$ respectively and X is a convex set. Then the discrete CCCP (Concave Convex Procedure Algorithm) algorithm [2]

$$\boldsymbol{\xi}_{t+1} \in \mathcal{A}(\boldsymbol{\xi}_t), t = 0, 1, \cdots \tag{26}$$

given by:

$$\nabla_{\boldsymbol{\xi}}\mathrm{E}_1\left(\boldsymbol{\xi}_{t+1}\right) = \nabla_{\boldsymbol{\xi}}\mathrm{E}_2(\boldsymbol{\xi}_t) \tag{27}$$

guarantees $\mathcal{A}$ to be monotonic with respect to E.
In addition, if E_1 or E_2 is strictly convex function, and $\boldsymbol{\xi}_{t+1} \neq \boldsymbol{\xi}_t$, then

$$\mathrm{E}\left(\boldsymbol{\xi}_{t+1}\right) < \mathrm{E}(\boldsymbol{\xi}_t) \tag{28}$$

Proof. E_1 and E_2 are convex functions, so that $\forall t \geqslant 0$,

$$
\left.\begin{array}{l}
\mathrm{E}_1\left(\boldsymbol{\xi}_t\right) \geqslant \mathrm{E}_1\left(\boldsymbol{\xi}_{t+1}\right)+\left(\nabla \mathrm{E}_1\left(\boldsymbol{\xi}_{t+1}\right)\right)^T\left(\boldsymbol{\xi}_t-\boldsymbol{\xi}_{t+1}\right) \\
\mathrm{E}_2\left(\boldsymbol{\xi}_{t+1}\right) \geqslant \mathrm{E}_2\left(\boldsymbol{\xi}_t\right)+\left(\nabla \mathrm{E}_2\left(\boldsymbol{\xi}_t\right)\right)^T\left(\boldsymbol{\xi}_{t+1}-\boldsymbol{\xi}_t\right) \\
\nabla \mathrm{E}_1\left(\boldsymbol{\xi}_{t+1}\right)=\nabla \mathrm{E}_2\left(\boldsymbol{\xi}_t\right)
\end{array}\right\} \\
\Longrightarrow \mathrm{E}_1\left(\boldsymbol{\xi}_t\right)+\mathrm{E}_2\left(\boldsymbol{\xi}_{t+1}\right) \geqslant \mathrm{E}_1\left(\boldsymbol{\xi}_{t+1}\right)+\mathrm{E}_2\left(\boldsymbol{\xi}_t\right) \\
\Longrightarrow \mathrm{E}_1\left(\boldsymbol{\xi}_{t+1}\right)-\mathrm{E}_2\left(\boldsymbol{\xi}_{t+1}\right) \leqslant \mathrm{E}_1\left(\boldsymbol{\xi}_t\right)-\mathrm{E}_2\left(\boldsymbol{\xi}_t\right) \\
\Longrightarrow \mathrm{E}\left(\boldsymbol{\xi}_{t+1}\right) \leqslant \mathrm{E}\left(\boldsymbol{\xi}_t\right) \tag{29}
$$

Therefore $\mathcal{A}$ is monotonic with respect to E.
In addition, if E_1 is strictly convex function, and $\boldsymbol{\xi}_{t+1} \neq \boldsymbol{\xi}_t$, then

$$
\left.\begin{array}{l}
\mathrm{E}_1\left(\boldsymbol{\xi}_t\right) > \mathrm{E}_1\left(\boldsymbol{\xi}_{t+1}\right)+\left(\nabla \mathrm{E}_1\left(\boldsymbol{\xi}_{t+1}\right)\right)^T\left(\boldsymbol{\xi}_t-\boldsymbol{\xi}_{t+1}\right) \\
\mathrm{E}_2\left(\boldsymbol{\xi}_{t+1}\right) \geqslant \mathrm{E}_2\left(\boldsymbol{\xi}_t\right)+\left(\nabla \mathrm{E}_2\left(\boldsymbol{\xi}_t\right)\right)^T\left(\boldsymbol{\xi}_{t+1}-\boldsymbol{\xi}_t\right) \\
\nabla \mathrm{E}_1\left(\boldsymbol{\xi}_{t+1}\right)=\nabla \mathrm{E}_2\left(\boldsymbol{\xi}_t\right)
\end{array}\right\} \\
\Longrightarrow \mathrm{E}_1\left(\boldsymbol{\xi}_t\right)+\mathrm{E}_2\left(\boldsymbol{\xi}_{t+1}\right) > \mathrm{E}_1\left(\boldsymbol{\xi}_{t+1}\right)+\mathrm{E}_2\left(\boldsymbol{\xi}_t\right) \\
\Longrightarrow \mathrm{E}_1\left(\boldsymbol{\xi}_{t+1}\right)-\mathrm{E}_2\left(\boldsymbol{\xi}_{t+1}\right) < \mathrm{E}_1\left(\boldsymbol{\xi}_t\right)-\mathrm{E}_2\left(\boldsymbol{\xi}_t\right) \\
\Longrightarrow \mathrm{E}\left(\boldsymbol{\xi}_{t+1}\right) < \mathrm{E}\left(\boldsymbol{\xi}_t\right) \tag{30}
$$

□

13 Lemma 6 Convex Concave Decomposition

The energy function $\mathrm{E}(\boldsymbol{\xi})$ is the difference of a strictly convex function $\mathrm{E}_1(\boldsymbol{\xi})$ and a convex function $\mathrm{E}_2(\boldsymbol{\xi})$:

$$
\begin{aligned}
\mathrm{E}(\boldsymbol{\xi}) &= \mathrm{E}_1(\boldsymbol{\xi})-\mathrm{E}_2(\boldsymbol{\xi}) \\
\mathrm{E}_1(\boldsymbol{\xi}) &= \frac{1}{2}\boldsymbol{\xi}^T\boldsymbol{\xi}+\beta^{-1}\ln N+\frac{1}{2}M^2=\frac{1}{2}\boldsymbol{\xi}^T\boldsymbol{\xi}+C_1 \\
\mathrm{E}_2(\boldsymbol{\xi}) &= \mathrm{lse}\left(\beta, \boldsymbol{X}^T\boldsymbol{\xi}\right)
\end{aligned} \tag{31}
$$

where C_1 does not depend on $\boldsymbol{\xi}$.

Proof. $\frac{1}{2}\boldsymbol{\xi}^T\boldsymbol{\xi}$ is a strictly convex function.
According to Lemma 4, lse is a convex function.

□

14 Lemma 7 Constraint Set

Let

$$S = \{\boldsymbol{\xi} | \|\boldsymbol{\xi}\|_2 \leqslant M\} \tag{32}$$

which is a convex and compact set.
Then

$$\forall t \geqslant 0, \boldsymbol{\xi}_{t+1} \in S \tag{33}$$

where $\boldsymbol{\xi}_{t+1}$ is from update rule 23.

Proof.

$$\left\|\boldsymbol{\xi}_{t+1}\right\|_2 = \|\boldsymbol{X}\mathbf{p}\|_2 = \left\|\sum_{i=1}^{N} p_i \mathbf{x}_i\right\|_2 \leqslant \sum_{i=1}^{N} p_i \left\|\mathbf{x}_i\right\|_2 \leqslant \sum_{i=1}^{N} p_i M = M \tag{34}$$

□

15 Lemma 8 Constraint Function

Define function

$$c(\boldsymbol{\xi}) = \boldsymbol{\xi}^T \boldsymbol{\xi} - M \tag{35}$$

Then

$$S = \{\boldsymbol{\xi} : c(\boldsymbol{\xi}) \leqslant 0\} \tag{36}$$

where S is defined by 32.

Proof.

$$\forall \boldsymbol{\xi}, c(\boldsymbol{\xi}) \leqslant 0 \iff \boldsymbol{\xi} \in S \tag{37}$$

□

16 Lemma 9 Convergence Theorem 1

Let a point-to-set map $\mathcal{A} : X \to \mathscr{P}(X)$ be a point-to-set map (an algorithm) that given a point $\boldsymbol{\xi}_0 \in X$ generates a sequence $\{\boldsymbol{\xi}_t\}_{t=0}^{\infty}$ through the iteration $\boldsymbol{\xi}_{t+1} \in \mathcal{A}(\boldsymbol{\xi}_t)$. Suppose

1. All points $\boldsymbol{\xi}_{t+1}, t \geqslant 0$ are in a compact set $\Omega \subset X$.

2. $\mathcal{A}$ is monotonic with respect to a continuous function E.

Then $\mathrm{E}(\boldsymbol{\xi}_t) \to \mathrm{E}(\boldsymbol{\xi}^*)$ for $t \to \infty$ and a limit point $\boldsymbol{\xi}^*$.

Proof. All points $\boldsymbol{\xi}_{t+1}, t \geqslant 0$ are in a compact set $\Omega \subset X$. Hence $\{\boldsymbol{\xi}_t\}_{t=0}^{\infty}$ must have a convergent subsequence

$$\boldsymbol{\xi}_{t_k} \to \boldsymbol{\xi}^*, \text{ as } k \to \infty \tag{38}$$

where $\boldsymbol{\xi}^* \in \Omega$ is a limit point. Continuity of E provides

$$\mathrm{E}\left(\boldsymbol{\xi}_{t_k}\right) \to \mathrm{E}(\boldsymbol{\xi}^*), \text{ as } k \to \infty \tag{39}$$

$\mathcal{A}$ is monotonic with respect to E. Hence

$$\forall t \geqslant 0, \mathrm{E}(\boldsymbol{\xi}_t) \geqslant \mathrm{E}(\boldsymbol{\xi}^*) \tag{40}$$

Using 39 and the definition of limit, given $\varepsilon_1 > 0$, there is a t_{ε_1} such that

$$\mathrm{E}\left(\boldsymbol{\xi}_{t_{\varepsilon_1}}\right) < \mathrm{E}(\boldsymbol{\xi}^*) + \varepsilon_1 \tag{41}$$

$\mathcal{A}$ is monotonic with respect to E, then

$$\forall t > t_{\varepsilon_1}, \mathrm{E}(\boldsymbol{\xi}_t) \leqslant \mathrm{E}\left(\boldsymbol{\xi}_{t_{\varepsilon_1}}\right) < \mathrm{E}(\boldsymbol{\xi}^*) + \varepsilon_1 \tag{42}$$

40, 42 then yield

$$\forall t > t_{\varepsilon_1}, |\mathrm{E}(\boldsymbol{\xi}_t) - \mathrm{E}(\boldsymbol{\xi}^*)| < \varepsilon_1 \tag{43}$$

Therefore

$$\mathrm{E}(\boldsymbol{\xi}_t) \to \mathrm{E}(\boldsymbol{\xi}^*), \text{ as } t \to \infty \tag{44}$$

□

17 Theorem 1 Global Convergence: Energy

The update rule 23 converges globally: For $\boldsymbol{\xi}_{t+1} = \mathcal{F}(\boldsymbol{\xi}_t)$, the energy $\mathrm{E}(\boldsymbol{\xi}_t) \to \mathrm{E}(\boldsymbol{\xi}^*)$ for $t \to \infty$ and a limit point $\boldsymbol{\xi}^*$.

Proof. Define minimization problem:

$$\min_{\boldsymbol{\xi}} \mathrm{E}(\boldsymbol{\xi}) = \min_{\boldsymbol{\xi}} \left(\mathrm{E}_1(\boldsymbol{\xi}) - \mathrm{E}_2(\boldsymbol{\xi})\right) \tag{45}$$

where $\mathrm{E}_1, \mathrm{E}_2$ are defined by 31.

1. **Applying CCCP Algorithm**

   According to 27, let

   $$\nabla_{\boldsymbol{\xi}} \mathrm{E}_1\left(\boldsymbol{\xi}_{t+1}\right) = \nabla_{\boldsymbol{\xi}} \mathrm{E}_2\left(\boldsymbol{\xi}_t\right) \tag{46}$$

   Using 4, 31, the resulting update rule is:

   $$\boldsymbol{\xi}_{t+1} = \boldsymbol{X} \operatorname{softmax}\left(\beta \boldsymbol{X}^T \boldsymbol{\xi}_t\right) \tag{47}$$

   which is equivalent to 23.

   Let

   $$\mathcal{A}(\boldsymbol{\xi}_t) = \left\{\boldsymbol{X} \operatorname{softmax}\left(\beta \boldsymbol{X}^T \boldsymbol{\xi}_t\right)\right\} \tag{48}$$

2. **Energy Convergence**

   Since

   (a) According to 32, 33, all points $\boldsymbol{\xi}_{t+1}, t \geqslant 0$ are in a compact set S.

   (b) According to Lemma 5, $\mathcal{A}$ is monotonic with respect to E.

   (c) Obviously E is continuous.

   then using Lemma 9

   $$\mathrm{E}(\boldsymbol{\xi}_t) \to \mathrm{E}(\boldsymbol{\xi}^*), \text{ as } t \to \infty \tag{49}$$

□

18 Lemma 10 MM Algorithm

Consider the optimization function $\mathrm{E} : X \to \mathbb{R}$:

$$\mathrm{E}(\boldsymbol{\xi}) = \mathrm{E}_1(\boldsymbol{\xi}) - \mathrm{E}_2(\boldsymbol{\xi}), \boldsymbol{\xi} \in \Omega \tag{50}$$

where E_1 and E_2 are both convex, E_2 is differentiable, and $\Omega \subset X$ is a convex set. Let

$$g(\boldsymbol{\xi}, \boldsymbol{\xi}_t) = \mathrm{E}_1(\boldsymbol{\xi}) - \mathrm{E}_2(\boldsymbol{\xi}_t) - (\nabla_{\boldsymbol{\xi}} \mathrm{E}_2(\boldsymbol{\xi}_t))^T (\boldsymbol{\xi} - \boldsymbol{\xi}_t) \tag{51}$$

Then the MM (Majorization-Minimization) algorithm [3]

$$\begin{aligned} \boldsymbol{\xi}_{t+1} \in \mathcal{A}(\boldsymbol{\xi}_t) &= \arg\min_{\boldsymbol{\xi} \in \Omega} g(\boldsymbol{\xi}, \boldsymbol{\xi}_t) \\ &= \arg\min_{\boldsymbol{\xi} \in \Omega} \Big(\mathrm{E}_1(\boldsymbol{\xi}) - (\nabla_{\boldsymbol{\xi}} \mathrm{E}_2(\boldsymbol{\xi}_t))^T \boldsymbol{\xi} \Big) \\ t = 0, 1, \cdots & \end{aligned} \tag{52}$$

guarantees $\mathcal{A}$ to be monotonic with respect to E.
In addition, if E_1 is strictly convex, then $\mathcal{A}$ is strictly monotonic with respect to E.

Proof. Since E_2 is convex and differentiable, then the first order characterization of convexity holds:

$$\mathrm{E}_2(\boldsymbol{\xi}) \geqslant \mathrm{E}_2(\boldsymbol{\xi}_t) + (\nabla_{\boldsymbol{\xi}} \mathrm{E}_2(\boldsymbol{\xi}_t))^T (\boldsymbol{\xi} - \boldsymbol{\xi}_t) \tag{53}$$

Therefore:

$$\begin{aligned} \mathrm{E}(\boldsymbol{\xi}) &= \mathrm{E}_1(\boldsymbol{\xi}) - \mathrm{E}_2(\boldsymbol{\xi}) \\ &\leqslant \mathrm{E}_1(\boldsymbol{\xi}) - \mathrm{E}_2(\boldsymbol{\xi}_t) - (\nabla_{\boldsymbol{\xi}} \mathrm{E}_2(\boldsymbol{\xi}_t))^T (\boldsymbol{\xi} - \boldsymbol{\xi}_t) \\ &= g(\boldsymbol{\xi}, \boldsymbol{\xi}_t) \end{aligned} \tag{54}$$

According to 52

$$\forall \boldsymbol{\xi}_{t+1} \in \mathcal{A}(\boldsymbol{\xi}_t), g\left(\boldsymbol{\xi}_{t+1}, \boldsymbol{\xi}_t\right) \leqslant g(\boldsymbol{\xi}_t, \boldsymbol{\xi}_t) \tag{55}$$

Using 51, 54, 55

$$\forall \boldsymbol{\xi}_{t+1} \in \mathcal{A}(\boldsymbol{\xi}_t), \mathrm{E}\left(\boldsymbol{\xi}_{t+1}\right) \leqslant g\left(\boldsymbol{\xi}_{t+1}, \boldsymbol{\xi}_t\right) \leqslant g(\boldsymbol{\xi}_t, \boldsymbol{\xi}_t) = \mathrm{E}(\boldsymbol{\xi}_t) \tag{56}$$

Thus

$$\forall \boldsymbol{\xi}_{t+1} \in \mathcal{A}\left(\boldsymbol{\xi}_t\right), \mathrm{E}\left(\boldsymbol{\xi}_{t+1}\right) \leqslant \mathrm{E}\left(\boldsymbol{\xi}_t\right) \tag{57}$$

Therefore $\mathcal{A}$ is monotonic with respect to E.
If E_1 is strictly convex with respect to $\boldsymbol{\xi}$, then $g\left(\boldsymbol{\xi}, \boldsymbol{\xi}_t\right)$ is also strictly convex with respect to $\boldsymbol{\xi}$ according to its definition 51. Since Ω is convex, then there exists only one minimum in 52, which is the global minimum.
Hence

$$\forall \boldsymbol{\xi}_{t+1} \in \mathcal{A}\left(\boldsymbol{\xi}_t\right), \boldsymbol{\xi}_{t+1} \neq \boldsymbol{\xi}_t \implies g\left(\boldsymbol{\xi}_{t+1}, \boldsymbol{\xi}_t\right) < g\left(\boldsymbol{\xi}_t, \boldsymbol{\xi}_t\right) \tag{58}$$

Using 51, 54, 58

$$\begin{aligned} &\forall \boldsymbol{\xi}_{t+1} \in \mathcal{A}\left(\boldsymbol{\xi}_t\right), \boldsymbol{\xi}_{t+1} \neq \boldsymbol{\xi}_t \implies \\ &\mathrm{E}\left(\boldsymbol{\xi}_{t+1}\right) \leqslant g\left(\boldsymbol{\xi}_{t+1}, \boldsymbol{\xi}_t\right) < g\left(\boldsymbol{\xi}_t, \boldsymbol{\xi}_t\right) = \mathrm{E}\left(\boldsymbol{\xi}_t\right) \end{aligned} \tag{59}$$

That is

$$\forall \boldsymbol{\xi}_{t+1} \in \mathcal{A}\left(\boldsymbol{\xi}_t\right), \boldsymbol{\xi}_{t+1} \neq \boldsymbol{\xi}_t \implies \mathrm{E}\left(\boldsymbol{\xi}_{t+1}\right) < \mathrm{E}\left(\boldsymbol{\xi}_t\right) \tag{60}$$

If $\mathcal{A}\left(\boldsymbol{\xi}_t\right) \neq \left\{\boldsymbol{\xi}_t\right\}$, suppose $\boldsymbol{\xi}_t \in \mathcal{A}\left(\boldsymbol{\xi}_t\right)$, then $\exists \boldsymbol{\xi}'_{t+1} \in \mathcal{A}\left(\boldsymbol{\xi}_t\right), \boldsymbol{\xi}'_{t+1} \neq \boldsymbol{\xi}_t$. Using 58

$$g\left(\boldsymbol{\xi}'_{t+1}, \boldsymbol{\xi}_t\right) < g\left(\boldsymbol{\xi}_t, \boldsymbol{\xi}_t\right) \tag{61}$$

Hence

$$\boldsymbol{\xi}_{t+1} \notin \mathcal{A}\left(\boldsymbol{\xi}_t\right) = \underset{\boldsymbol{\xi} \in \Omega}{\arg\min} g\left(\boldsymbol{\xi}, \boldsymbol{\xi}_t\right) \tag{62}$$

which is a contradiction. Thus if $\mathcal{A}\left(\boldsymbol{\xi}_t\right) \neq \left\{\boldsymbol{\xi}_t\right\}$, then $\boldsymbol{\xi}_{t+1} \notin \mathcal{A}\left(\boldsymbol{\xi}_t\right)$. That is $\forall \boldsymbol{\xi}_{t+1} \in \mathcal{A}\left(\boldsymbol{\xi}_t\right), \boldsymbol{\xi}_{t+1} \neq \boldsymbol{\xi}_t$. Using 60, $\mathrm{E}\left(\boldsymbol{\xi}_{t+1}\right) < \mathrm{E}\left(\boldsymbol{\xi}_t\right)$. It follows that

$$\mathcal{A}\left(\boldsymbol{\xi}_t\right) \neq \left\{\boldsymbol{\xi}_t\right\} \implies \forall \boldsymbol{\xi}_{t+1} \in \mathcal{A}\left(\boldsymbol{\xi}_t\right), \mathrm{E}\left(\boldsymbol{\xi}_{t+1}\right) < \mathrm{E}\left(\boldsymbol{\xi}_t\right) \tag{63}$$

Therefore point-set-map $\mathcal{A}$ is strictly monotonic with respect to E.

□

19 Lemma 11 Hopfield MM Algorithm Update Rule

$$\arg\min_{\boldsymbol{\xi}} g\left(\boldsymbol{\xi}, \boldsymbol{\xi}_t\right) = \arg\min_{\boldsymbol{\xi} \in S} g\left(\boldsymbol{\xi}, \boldsymbol{\xi}_t\right) = \left\{ \boldsymbol{X} \operatorname{softmax}\left(\beta \boldsymbol{X}^T \boldsymbol{\xi}_t\right) \right\} \tag{64}$$

where g is defined by 51, $\mathrm{E}_1, \mathrm{E}_2$ are defined by 31, S is a convex and compact set defined by 32.

Proof. E_1 is strictly convex with respect to $\boldsymbol{\xi}$, then $g\left(\boldsymbol{\xi}, \boldsymbol{\xi}_t\right)$ is also strictly convex with respect to $\boldsymbol{\xi}$ according to its definition 51. Then there exists only one minimum in 52, which is the global minimum.
Let

$$\nabla_{\boldsymbol{\xi}} g\left(\boldsymbol{\xi}, \boldsymbol{\xi}_t\right) = \mathbf{0} \tag{65}$$

Then

$$\nabla_{\boldsymbol{\xi}} \mathrm{E}_1\left(\boldsymbol{\xi}\right) = \nabla_{\boldsymbol{\xi}} \mathrm{E}_2\left(\boldsymbol{\xi}_t\right) \tag{66}$$

Using 4, 31

$$\boldsymbol{\xi} = \boldsymbol{X} \operatorname{softmax}\left(\beta \boldsymbol{X}^T \boldsymbol{\xi}_t\right) \tag{67}$$

Using 33

$$\boldsymbol{\xi} \in S \tag{68}$$

Therefore

$$\arg\min_{\boldsymbol{\xi}} g\left(\boldsymbol{\xi}, \boldsymbol{\xi}_t\right) = \arg\min_{\boldsymbol{\xi} \in S} g\left(\boldsymbol{\xi}, \boldsymbol{\xi}_t\right) = \left\{ \boldsymbol{X} \operatorname{softmax}\left(\beta \boldsymbol{X}^T \boldsymbol{\xi}_t\right) \right\} \tag{69}$$

□

20 Lemma 12 Closedness Sufficient Condition

Given a continuous function $h\left(\mathbf{x}, \mathbf{y}\right)$ on $X \times Y$, where X and Y are closed sets, define the point-to-set map $\mathcal{A} : X \to \mathscr{P}(Y)$ by

$$\mathcal{A}\left(\mathbf{x}\right) = \arg\min_{\mathbf{y} \in Y} h\left(\mathbf{x}, \mathbf{y}\right) \tag{70}$$

If $\mathcal{A}$ is nonempty at each $\mathbf{x} \in X$, then $\mathcal{A}$ is closed on X [4].

Proof. According to closedness definition, suppose

$$\begin{cases} \mathbf{x}_t \to \mathbf{x}_0 \text{ as } t \to \infty \\ \mathbf{y}_t \to \mathbf{y}_0 \text{ as } t \to \infty \\ \mathbf{x}_t \in X \\ \mathbf{y}_t \in \mathcal{A}\left(\mathbf{x}_t\right) = \underset{\mathbf{y} \in Y}{\arg\min} h\left(\mathbf{x}_t, \mathbf{y}\right) \end{cases} \tag{71}$$

Since X and Y are closed sets, $(\mathbf{x}_0, \mathbf{y}_0) \in X \times Y$.
By continuity of h

$$\begin{gathered} h\left(\mathbf{x}_t, \mathbf{y}_t\right) \to h\left(\mathbf{x}_0, \mathbf{y}_0\right) \text{ as } t \to \infty \\ \forall \mathbf{y} \in Y, h\left(\mathbf{x}_t, \mathbf{y}\right) \to h\left(\mathbf{x}_0, \mathbf{y}\right) \text{ as } t \to \infty \end{gathered} \tag{72}$$

Using the definition of limit, $\forall \mathbf{y} \in Y, \forall \varepsilon_2 > 0$, there is a t_{ε_2} such that $\forall t > t_{\varepsilon_2}$,

$$\begin{gathered} h\left(\mathbf{x}_t, \mathbf{y}_t\right) > h\left(\mathbf{x}_0, \mathbf{y}_0\right) - \varepsilon_2 \\ h\left(\mathbf{x}_t, \mathbf{y}\right) < h\left(\mathbf{x}_0, \mathbf{y}\right) + \varepsilon_2 \end{gathered} \tag{73}$$

and

$$\mathbf{y}_t \in \mathcal{A}\left(\mathbf{x}_t\right) = \underset{\mathbf{y} \in Y}{\arg\min} h\left(\mathbf{x}_t, \mathbf{y}\right) \implies h\left(\mathbf{x}_t, \mathbf{y}_t\right) \leqslant h\left(\mathbf{x}_t, \mathbf{y}\right) \tag{74}$$

Using 73, 74

$$h\left(\mathbf{x}_0, \mathbf{y}_0\right) < h\left(\mathbf{x}_t, \mathbf{y}_t\right) + \varepsilon_2 \leqslant h\left(\mathbf{x}_t, \mathbf{y}\right) + \varepsilon_2 < h\left(\mathbf{x}_0, \mathbf{y}\right) + 2\varepsilon_2 \tag{75}$$

Hence

$$h\left(\mathbf{x}_0, \mathbf{y}_0\right) \leqslant h\left(\mathbf{x}_0, \mathbf{y}\right) \tag{76}$$

It follows that

$$\mathbf{y}_0 \in \mathcal{A}\left(\mathbf{x}_0\right) = \underset{\mathbf{y} \in Y}{\arg\min} h\left(\mathbf{x}_0, \mathbf{y}\right) \tag{77}$$

Therefore $\mathcal{A}$ is closed on X.

□

21 Lemma 13 Convergence Theorem 2

Let a point-to-set map $\mathcal{A} : X \to \mathscr{P}(X)$ be a point-to-set map (an algorithm) that given a point $\boldsymbol{\xi}_0 \in X$ generates a sequence $\{\boldsymbol{\xi}_t\}_{t=0}^{\infty}$ through the iteration $\boldsymbol{\xi}_{t+1} \in \mathcal{A}(\boldsymbol{\xi}_t)$. Also let a solution set $\Gamma \subset X$ be given. Suppose

1. All points $\boldsymbol{\xi}_{t+1}, t \geqslant 0$ are in a compact set $\Omega \subset X$.
2. There is a continuous function $\mathrm{E} : X \to \mathbb{R}$ such that:
   (a) $\boldsymbol{\xi}_t \notin \Gamma \implies \mathrm{E}\left(\boldsymbol{\xi}_{t+1}\right) < \mathrm{E}(\boldsymbol{\xi}_t), \forall \boldsymbol{\xi}_{t+1} \in \mathcal{A}(\boldsymbol{\xi}_t)$
   (b) $\boldsymbol{\xi}_t \in \Gamma \implies \mathrm{E}\left(\boldsymbol{\xi}_{t+1}\right) \leqslant \mathrm{E}(\boldsymbol{\xi}_t), \forall \boldsymbol{\xi}_{t+1} \in \mathcal{A}(\boldsymbol{\xi}_t)$
3. $\mathcal{A}$ is closed on X.

Then all limit points of the sequence $\{\boldsymbol{\xi}_t\}_{t=0}^{\infty}$ are in Γ [5].

Proof. Given any convergent subsequence of $\{\boldsymbol{\xi}_t\}_{t=0}^{\infty}$

$$\boldsymbol{\xi}_{t_k} \to \boldsymbol{\xi}^*, \text{ as } k \to \infty \tag{78}$$

where $\boldsymbol{\xi}^* \in \Omega$ is a limit point, consider subsequence:

$$\left\{\boldsymbol{\xi}_{t_k+1}\right\}_{k=0}^{\infty} \tag{79}$$

Since all points $\boldsymbol{\xi}_{t+1}, t \geqslant 0$ are in the compact set Ω, then $\left\{\boldsymbol{\xi}_{t_k+1}\right\}_{k=0}^{\infty}$ must have a convergent subsequence

$$\boldsymbol{\xi}_{t_{k_l}+1} \to \boldsymbol{\xi}^{**}, \text{ as } l \to \infty \tag{80}$$

where $\boldsymbol{\xi}^{**} \in \Omega$ is a limit point.
Using 78

$$\boldsymbol{\xi}_{t_{k_l}} \to \boldsymbol{\xi}^*, \text{ as } l \to \infty \tag{81}$$

It is known that

$$\boldsymbol{\xi}_{t_{k_l}+1} \in \mathcal{A}\left(\boldsymbol{\xi}_{t_{k_l}}\right) \tag{82}$$

Since $\mathcal{A}$ is closed on X, then using 80, 81 and 82

$$\boldsymbol{\xi}^{**} \in \mathcal{A}(\boldsymbol{\xi}^*) \tag{83}$$

Since

1. All points $\boldsymbol{\xi}_{t+1}, t \geqslant 0$ are in the compact set Ω.

2. $\mathcal{A}$ is monotonic with respect to a continuous function E.

According to Lemma 9,

$$\lim_{t \to \infty} \mathrm{E}\left(\boldsymbol{\xi}_t\right) = \mathrm{E}\left(\boldsymbol{\xi}^{**}\right) = \mathrm{E}\left(\boldsymbol{\xi}^{*}\right) \tag{84}$$

If $\boldsymbol{\xi}^* \notin \Gamma$, then $\mathrm{E}\left(\boldsymbol{\xi}^{**}\right) < \mathrm{E}\left(\boldsymbol{\xi}^{*}\right)$. A contradiction. Therefore $\boldsymbol{\xi}^* \in \Gamma$.

□

22 Lemma 14 Fixed Points Sufficient Condition

Let a point-to-set map $\mathcal{A} : X \to \mathscr{P}(X)$ be a point-to-set map (an algorithm) that given a point $\boldsymbol{\xi}_0 \in X$ generates a sequence $\{\boldsymbol{\xi}_t\}_{t=0}^{\infty}$ through the iteration $\boldsymbol{\xi}_{t+1} \in \mathcal{A}\left(\boldsymbol{\xi}_t\right)$. Suppose

1. All points $\boldsymbol{\xi}_{t+1}, t \geqslant 0$ are in a compact set $\Omega \subset X$.

2. $\mathcal{A}$ is strictly monotonic with respect to a continuous function E.

3. $\mathcal{A}$ is closed on X.

Then all limit points of the sequence $\{\boldsymbol{\xi}_t\}_{t=0}^{\infty}$ are fixed points of $\mathcal{A}$.

Proof. Let

$$\Gamma = \{\boldsymbol{\xi} \in X | \mathcal{A}\left(\boldsymbol{\xi}\right) = \{\boldsymbol{\xi}\}\} \tag{85}$$

Using Lemma 13, all limit points of the sequence $\{\boldsymbol{\xi}_t\}_{t=0}^{\infty}$ are in Γ.
Therefore all limit points are fixed points of $\mathcal{A}$.

□

23 Lemma 15 Stationary Points Sufficient Condition

Suppose $\mathrm{E}_1\left(\boldsymbol{\xi}\right), \mathrm{E}_2\left(\boldsymbol{\xi}\right)$ are differentiable.
And

$$\Omega = \{\boldsymbol{\xi} : c_i\left(\boldsymbol{\xi}\right) \leqslant 0, i \in [m], d_j\left(\boldsymbol{\xi}\right) = 0, j \in [p]\} \tag{86}$$

$\boldsymbol{\xi}^*$ is a generalized fixed point of 52, then $\boldsymbol{\xi}^*$ is a stationary point of the the program in 50.

Proof. We have $\boldsymbol{\xi}^* \in \mathcal{A}(\boldsymbol{\xi}^*)$ and $\boldsymbol{\xi}^* \in \Omega$. Then there exist Lagrange multipliers $\{\eta_i^*\}_{i=1}^m \subset \mathbb{R}^+$ and $\left\{\mu_j^*\right\}_{j=1}^p \subset \mathbb{R}$ such that the following KKT conditions hold:

$$\begin{cases} \nabla_{\boldsymbol{\xi}} \mathrm{E}_1(\boldsymbol{\xi}^*) - \nabla_{\boldsymbol{\xi}} \mathrm{E}_2(\boldsymbol{\xi}^*) + \sum_{i=1}^m \eta_i^* \nabla_{\boldsymbol{\xi}} c_i(\boldsymbol{\xi}^*) + \sum_{j=1}^p \mu_j^* \nabla_{\boldsymbol{\xi}} d_j(\boldsymbol{\xi}^*) = \mathbf{0} \\ c_i(\boldsymbol{\xi}^*) \leqslant 0, \eta_i^* \geqslant 0, c_i(\boldsymbol{\xi}^*)\, \eta_i^* = 0, \forall i \in [m] \\ d_j(\boldsymbol{\xi}^*) = 0, \mu_j^* \in \mathbb{R}, \forall j \in [p] \end{cases} \tag{87}$$

is exactly KKT conditions of 50 which are satisfied by $\left(\boldsymbol{\xi}^*, \{\eta_i^*\}_{i=1}^m, \left\{\mu_j^*\right\}_{j=1}^p\right)$ and therefore, is a stationary point of 50.

□

24 Lemma 16 Convergence of Adjacent Point Differences

Let a point-to-set map $\mathcal{A} : X \to \mathscr{P}(X)$ be a point-to-set map (an algorithm) that given a point $\boldsymbol{\xi}_0 \in X$ generates a sequence $\{\boldsymbol{\xi}_t\}_{t=0}^\infty$ through the iteration $\boldsymbol{\xi}_{t+1} \in \mathcal{A}(\boldsymbol{\xi}_t)$. Suppose [6]

1. All points $\boldsymbol{\xi}_{t+1}, t \geqslant 0$ are in a compact set $\Omega \subset X$.
2. $\mathcal{A}$ is strictly monotonic with respect to a continuous function E.
3. $\mathcal{A}$ is closed on X.

Then

$$\left\|\boldsymbol{\xi}_{t+1} - \boldsymbol{\xi}_t\right\|_2 \to 0, \text{ as } t \to \infty \tag{88}$$

Proof. Suppose

$$\left\|\boldsymbol{\xi}_{t+1} - \boldsymbol{\xi}_t\right\|_2 \not\to 0, \text{ as } t \to \infty \tag{89}$$

Then there exists a $\varepsilon_3 > 0$ and a subsequence

$$\left\{\boldsymbol{\xi}_{t_k}\right\}_{k=0}^\infty \tag{90}$$

such that

$$\left\|\boldsymbol{\xi}_{t_k+1}-\boldsymbol{\xi}_{t_k}\right\|_2 \geqslant \varepsilon_3, \forall k \geqslant 0 \tag{91}$$

Since all points $\boldsymbol{\xi}_{t+1}, t \geqslant 0$ are in a compact set Ω, then $\left\{\boldsymbol{\xi}_{t_k}\right\}_{k=0}^{\infty}$ must have a convergent subsequence

$$\boldsymbol{\xi}_{t_{k_l}} \rightarrow \boldsymbol{\xi}^*, \text{ as } l \rightarrow \infty \tag{92}$$

where $\boldsymbol{\xi}^* \in \Omega$ is a limit point.
Consider subsequence:

$$\left\{\boldsymbol{\xi}_{t_{k_l}+1}\right\}_{k=0}^{\infty} \tag{93}$$

It must have a convergent subsequence

$$\boldsymbol{\xi}_{t_{k_{l_m}}+1} \rightarrow \boldsymbol{\xi}^{**}, \text{ as } m \rightarrow \infty \tag{94}$$

where $\boldsymbol{\xi}^{**} \in \Omega$ is a limit point.
Using 92

$$\boldsymbol{\xi}_{t_{k_{l_m}}} \rightarrow \boldsymbol{\xi}^*, \text{ as } m \rightarrow \infty \tag{95}$$

It is known that

$$\boldsymbol{\xi}_{t_{k_{l_m}}+1} \in \mathcal{A}\left(\boldsymbol{\xi}_{t_{k_{l_m}}}\right) \tag{96}$$

Since $\mathcal{A}$ is closed on X, using 94, 95 and 96

$$\boldsymbol{\xi}^{**} \in \mathcal{A}\left(\boldsymbol{\xi}^*\right) \tag{97}$$

Since

1. All points $\boldsymbol{\xi}_{t+1}, t \geqslant 0$ are in the compact set Ω.

2. $\mathcal{A}$ is monotonic with respect to a continuous function E.

According to Lemma 9,

$$\lim_{t \rightarrow \infty} \mathrm{E}\left(\boldsymbol{\xi}_t\right)=\mathrm{E}\left(\boldsymbol{\xi}^{**}\right)=\mathrm{E}\left(\boldsymbol{\xi}^*\right) \tag{98}$$

Since $\mathcal{A}$ is strictly monotonic with respect to E, using 97, 98

$$\mathcal{A}\left(\boldsymbol{\xi}^{*}\right)=\left\{\boldsymbol{\xi}^{*}\right\} \tag{99}$$

Hence

$$\boldsymbol{\xi}^{*}=\boldsymbol{\xi}^{**} \tag{100}$$

using 94, 95 and 100

$$\left\|\boldsymbol{\xi}_{t_{k_{l_m}}+1}-\boldsymbol{\xi}_{t_{k_{l_m}}}\right\|_{2} \rightarrow 0, \text{ as } m \rightarrow \infty \tag{101}$$

which is in contradiction with 91.
Therefore

$$\left\|\boldsymbol{\xi}_{t+1}-\boldsymbol{\xi}_{t}\right\|_{2} \rightarrow 0, \text{ as } t \rightarrow \infty \tag{102}$$

□

25 Lemma 17 Compactness of Limit Points

Let a point-to-set map $\mathcal{A}: X \rightarrow \mathscr{P}(X)$ be a point-to-set map (an algorithm) that given a point $\boldsymbol{\xi}_0 \in X$ generates a sequence $\{\boldsymbol{\xi}_t\}_{t=0}^{\infty}$ through the iteration $\boldsymbol{\xi}_{t+1} \in \mathcal{A}(\boldsymbol{\xi}_t)$.
Suppose all points $\boldsymbol{\xi}_{t+1}, t \geqslant 0$ are in a compact set $\Omega \subset X$.
Let S_{lim} denotes the set of limit points of $\{\boldsymbol{\xi}_t\}_{t=0}^{\infty}$.
Then S_{lim} is compact.

Proof. Since all points $\boldsymbol{\xi}_{t+1}, t \geqslant 0$ are in a compact set Ω, then S_{lim} is not empty and

$$\forall \boldsymbol{\xi}_{lim} \in S_{lim}, \boldsymbol{\xi}_{lim} \in \Omega \tag{103}$$

Thus S_{lim} is bounded.
For any convergent subsequence of S_{lim}

$$\left\{\boldsymbol{\xi}_{lim_k}\right\}_{k=0}^{\infty} \tag{104}$$

suppose

$$\boldsymbol{\xi}_{lim_k} \rightarrow \boldsymbol{\xi}_{lim}^{*} \text{ as } k \rightarrow \infty \tag{105}$$

Since $\forall k \geqslant 0, \boldsymbol{\xi}_{lim_k}$ is a limit point of $\{\boldsymbol{\xi}_t\}_{t=0}^{\infty}$, using the definition of limit,

$$\forall k \geqslant 0, \exists \boldsymbol{\xi}_{t_k} \in \{\boldsymbol{\xi}_t\}_{t=0}^{\infty}, \left\|\boldsymbol{\xi}_{t_k} - \boldsymbol{\xi}_{lim_k}\right\|_2 < \frac{1}{k+1} \tag{106}$$

Then

$$\left\|\boldsymbol{\xi}_{t_k} - \boldsymbol{\xi}_{lim_k}\right\|_2 \to 0, k \to \infty \tag{107}$$

That is

$$\boldsymbol{\xi}_{t_k} - \boldsymbol{\xi}_{lim_k} \to \mathbf{0}, k \to \infty \tag{108}$$

Using 105, 108

$$\boldsymbol{\xi}_{t_k} = \left(\boldsymbol{\xi}_{t_k} - \boldsymbol{\xi}_{lim_k}\right) + \boldsymbol{\xi}_{lim_k} \to \boldsymbol{\xi}_{lim}^*, k \to \infty \tag{109}$$

Since S_{lim} denotes the set of limit points of $\{\boldsymbol{\xi}_t\}_{t=0}^{\infty}$, it follows that

$$\boldsymbol{\xi}^* \in S_{lim} \tag{110}$$

Therefore S_{lim} is closed. It is also bounded, so it is a compact set.

□

26 Lemma 18 Connectedness of Limit Points

Let a point-to-set map $\mathcal{A} : X \to \mathscr{P}(X)$ be a point-to-set map (an algorithm) that given a point $\boldsymbol{\xi}_0 \in X$ generates a sequence $\{\boldsymbol{\xi}_t\}_{t=0}^{\infty}$ through the iteration $\boldsymbol{\xi}_{t+1} \in \mathcal{A}(\boldsymbol{\xi}_t)$.
Suppose

1. All points $\boldsymbol{\xi}_{t+1}, t \geqslant 0$ are in a compact set $\Omega \subset X$.
2. $\left\|\boldsymbol{\xi}_{t+1} - \boldsymbol{\xi}_t\right\|_2 \to 0,$ as $t \to \infty$
3. $\{\boldsymbol{\xi}_t\}_{t=0}^{\infty}$ does not converge.

Let S_{lim} denotes the set of limit points of $\{\boldsymbol{\xi}_t\}_{t=0}^{\infty}$.
Then S_{lim} is connected.

Proof. Since $\{\boldsymbol{\xi}_t\}_{t=0}^{\infty}$ does not converge, then S_{lim} must contain at least two points. Suppose S_{lim} is not connected. Then it can be decomposed into the union of two nonempty closed sets of points without common points.

1. **Define Two Subsets**

   Let

$$S_{lim} = S_{lim}^1 \cup S_{lim}^2 \tag{111}$$

   where $S_{lim}^1 \cup S_{lim}^2$ are both nonempty, closed and:

$$S_{lim}^1 \cap S_{lim}^2 = \varnothing \tag{112}$$

2. **Define Distance to Set**

   Let

$$d = \inf\left\{ \left\| \boldsymbol{\xi}_{lim}^1 - \boldsymbol{\xi}_{lim}^2 \right\|_2 \middle| \boldsymbol{\xi}_{lim}^1 \in S_{lim}^1, \boldsymbol{\xi}_{lim}^2 \in S_{lim}^2 \right\} \tag{113}$$

   Suppose $d = 0$, then there is a sequence $\left\{\boldsymbol{\xi}_{lim_k}^1\right\}_{k=1}^{\infty}$ that belongs to S_{lim}^1, and a sequence $\left\{\boldsymbol{\xi}_{lim_k}^2\right\}_{k=1}^{\infty}$ that belongs to S_{lim}^2, and

$$\boldsymbol{\xi}_{lim_k}^1 - \boldsymbol{\xi}_{lim_k}^2 \to \mathbf{0}, k \to \infty \tag{114}$$

   Since S_{lim}^1 and S_{lim}^2 are closed, then each of the sequences $\left\{\boldsymbol{\xi}_{lim_k}^1\right\}_{k=1}^{\infty}$ and $\left\{\boldsymbol{\xi}_{lim_k}^2\right\}_{k=1}^{\infty}$ has a convergent subsequence. Using 114, two subsequences converge to the same limit point. Then this limit point belongs to both closed set S_{lim}^1 and S_{lim}^2, which is a contradiction with 112. Thus $d > 0$.

   Since S_{lim}^1 and S_{lim}^2 are nonempty, then there exists two limit points $\boldsymbol{\xi}_{lim}^{*1} \in S_{lim}^1, \boldsymbol{\xi}_{lim}^{*2} \in S_{lim}^2$.

   Suppose

$$\begin{cases} \boldsymbol{\xi}_{t_k}^1 \to \boldsymbol{\xi}_{lim}^{*1}, \text{ as } k \to \infty \\ \boldsymbol{\xi}_{t_k}^2 \to \boldsymbol{\xi}_{lim}^{*2}, \text{ as } k \to \infty \end{cases} \tag{115}$$

   where $Q_1 = \left\{\boldsymbol{\xi}_{t_k}^1\right\}_{k=0}^{\infty}$ and $Q_2 = \left\{\boldsymbol{\xi}_{t_k}^2\right\}_{k=0}^{\infty}$ are subsequences of $\{\boldsymbol{\xi}_t\}_{t=0}^{\infty}$.

   Since

$$\left\| \boldsymbol{\xi}_{t+1} - \boldsymbol{\xi}_t \right\|_2 \to 0, \text{ as } t \to \infty \tag{116}$$

then

$$\exists t_d, \forall t > t_d, \left\|\boldsymbol{\xi}_{t+1} - \boldsymbol{\xi}_t\right\|_2 < \frac{d}{3} \tag{117}$$

Let

$$\begin{cases} d_1\left(\boldsymbol{\xi}\right) = \inf\left\{\left\|\boldsymbol{\xi} - \boldsymbol{\xi}^1_{lim}\right\|_2 \middle| \boldsymbol{\xi} \in \left\{\boldsymbol{\xi}_t\right\}_{t=0}^{\infty}, \boldsymbol{\xi}^1_{lim} \in S^1_{lim}\right\} \\ d_2\left(\boldsymbol{\xi}\right) = \inf\left\{\left\|\boldsymbol{\xi} - \boldsymbol{\xi}^2_{lim}\right\|_2 \middle| \boldsymbol{\xi} \in \left\{\boldsymbol{\xi}_t\right\}_{t=0}^{\infty}, \boldsymbol{\xi}^2_{lim} \in S^2_{lim}\right\} \end{cases} \tag{118}$$

Using 113, 118

$$\forall \boldsymbol{\xi} \in \left\{\boldsymbol{\xi}_t\right\}_{t=0}^{\infty}, d_1\left(\boldsymbol{\xi}\right) + d_2\left(\boldsymbol{\xi}\right) \geqslant d \tag{119}$$

3. **Create Subsequence**

   Now let's create a subsequence of $\left\{\boldsymbol{\xi}_t\right\}_{t=0}^{\infty}$ whose limit points are outside both S^1_{lim} and S^2_{lim}:

   For any arbitrarily large number K, Using 115, 118

$$\begin{cases} \exists m > \max\left\{K, t_d\right\}, d_1\left(\boldsymbol{\xi}_m\right) < \dfrac{d}{3} \\ \exists n > m, d_2\left(\boldsymbol{\xi}_n\right) < \dfrac{d}{3} \end{cases} \tag{120}$$

   Define index set

$$I = \left\{i \middle| m \leqslant i \leqslant n\right\} \tag{121}$$

   Using 119, 120

$$\begin{cases} d_2\left(\boldsymbol{\xi}_m\right) \geqslant \dfrac{2}{3}d \\ d_1\left(\boldsymbol{\xi}_n\right) \geqslant \dfrac{2}{3}d \end{cases} \tag{122}$$

   Using 120, 121, 122

$$\exists t_{k_1} = \min\left\{i \in I \middle| d_1\left(\boldsymbol{\xi}_i\right) \geqslant \frac{d}{3}\right\} \tag{123}$$

Obviously:

$$\begin{cases} t_{k_1} > m \\ d_1\left(\boldsymbol{\xi}_{t_{k_1}}\right) \geqslant \dfrac{d}{3} \\ d_1\left(\boldsymbol{\xi}_{t_{k_1}-1}\right) < \dfrac{d}{3} \end{cases} \tag{124}$$

Using 119, 124

$$d_2\left(\boldsymbol{\xi}_{t_{k_1}-1}\right) \geqslant \frac{2}{3}d \tag{125}$$

Using 117, 125

$$d_2\left(\boldsymbol{\xi}_{t_{k_1}}\right) \geqslant \frac{d}{3} \tag{126}$$

Using 124, 126

$$\left.\begin{array}{l} \boldsymbol{\xi}_{t_{k_1}} \notin S^1_{lim} \\ \boldsymbol{\xi}_{t_{k_1}} \notin S^2_{lim} \end{array}\right\} \Longrightarrow \boldsymbol{\xi}_{t_{k_1}} \notin S^1_{lim} \cup S^2_{lim} \tag{127}$$

Follow this way, there exists an infinite sequence of $k_1, k_2, \cdots$ for which 127 holds. Therefore all limit points of this sequence are outside both S^1_{lim} and S^2_{lim}, which is in contradiction with 111. Therefore S_{lim} is connected.

□

27 Theorem 2 Global Convergence: Stationary Points

For the iteration 23 we have

$$\mathrm{E}\left(\boldsymbol{\xi}_t\right) \to \mathrm{E}\left(\boldsymbol{\xi}^*\right) = \mathrm{E}^*, \text{ as } t \to \infty \tag{128}$$

for some stationary point $\boldsymbol{\xi}^*$. Furthermore

$$\left\|\boldsymbol{\xi}_{t+1}-\boldsymbol{\xi}_t\right\|_2 \rightarrow 0, \text{ as } t \rightarrow \infty \tag{129}$$

And either $\{\boldsymbol{\xi}_t\}_{t=0}^{\infty}$ converges, or, in the other case, the set of limit points of $\{\boldsymbol{\xi}_t\}_{t=0}^{\infty}$ is a connected and compact subset of $\mathcal{L}\left(\mathrm{E}^*\right)$, where $\mathcal{L}(a)=\{\boldsymbol{\xi} \in \mathcal{L} | \mathrm{E}(\boldsymbol{\xi})=a\}$ and $\mathcal{L}$ is the set of stationary points of the iteration 23. If $\mathcal{L}\left(\mathrm{E}^*\right)$ is finite, then any sequence $\{\boldsymbol{\xi}_t\}_{t=0}^{\infty}$ generated by the iteration 23 converges to some $\boldsymbol{\xi}^* \in \mathcal{L}\left(\mathrm{E}^*\right)$.

Proof. Define minimization problem:

$$\min_{\boldsymbol{\xi}} \mathrm{E}(\boldsymbol{\xi})=\min_{\boldsymbol{\xi}}\left(\mathrm{E}_1(\boldsymbol{\xi})-\mathrm{E}_2(\boldsymbol{\xi})\right) \tag{130}$$

where $\mathrm{E}_1, \mathrm{E}_2$ are defined by 31.

1. **Applying MM Algorithm**

   According to 51, 52, let

$$g\left(\boldsymbol{\xi}, \boldsymbol{\xi}_t\right)=\mathrm{E}_1(\boldsymbol{\xi})-\mathrm{E}_2\left(\boldsymbol{\xi}_t\right)-\left(\nabla_{\boldsymbol{\xi}} \mathrm{E}_2\left(\boldsymbol{\xi}_t\right)\right)^T\left(\boldsymbol{\xi}-\boldsymbol{\xi}_t\right) \tag{131}$$

$$\begin{aligned}
\boldsymbol{\xi}_{t+1} \in \mathcal{A}\left(\boldsymbol{\xi}_t\right) &= \arg\min_{\boldsymbol{\xi}} g\left(\boldsymbol{\xi}, \boldsymbol{\xi}_t\right) \\
&= \arg\min_{\boldsymbol{\xi}} g\Big(\mathrm{E}_1(\boldsymbol{\xi})-\left(\nabla_{\boldsymbol{\xi}} \mathrm{E}_2\left(\boldsymbol{\xi}_t\right)\right)^T \boldsymbol{\xi}\Big) \\
t=0,1,\cdots &
\end{aligned} \tag{132}$$

   According to 64, the resulting update rule is:

$$\boldsymbol{\xi}_{t+1}=\boldsymbol{X} \operatorname{softmax}\left(\beta \boldsymbol{X}^T \boldsymbol{\xi}_t\right) \tag{133}$$

   which is equivalent to 23.

2. **$\mathcal{A}$ is Strictly Monotonic**

   $\mathrm{E}_1(\boldsymbol{\xi})=\frac{1}{2} \boldsymbol{\xi}^T \boldsymbol{\xi}+C_1$ is strictly convex.

   $\mathrm{E}_2(\boldsymbol{\xi})$ is convex and differentiable using 4, 31.

   Then using Lemma 10, $\mathcal{A}$ is strictly monotonic with respect to E.

3. **$\mathcal{A}$ is Uniformly Compact**

   Using 64

   All points $\boldsymbol{\xi}_{t+1}, t \geqslant 0$ are in the compact set S using 64, therefore $\mathcal{A}$ is uniformly compact on $\mathbb{R}^d$.

4. **$\mathcal{A}$ is Closed**

   Let

$$h(\mathbf{x}, \mathbf{y}) = g(\mathbf{y}, \mathbf{x}) \tag{134}$$

   It follows that

$$\mathcal{A}(\boldsymbol{\xi}) = \arg\min_{\boldsymbol{\xi}} g(\boldsymbol{\xi}, \boldsymbol{\xi}_t) = \arg\min_{\boldsymbol{\xi}} h(\boldsymbol{\xi}_t, \boldsymbol{\xi}) \tag{135}$$

   Now let's prove $\mathcal{A}$ is closed on $\mathbb{R}^d$:

   (a) $\mathcal{A}(\boldsymbol{\xi}_t)$ is nonempty:
   $\mathrm{E}_1(\boldsymbol{\xi}) = \frac{1}{2}\boldsymbol{\xi}^T\boldsymbol{\xi} + C_1$ is continuous, thus given any $\boldsymbol{\xi}_t$, $g(\boldsymbol{\xi}, \boldsymbol{\xi}_t)$ is continuous at every $\boldsymbol{\xi}$ in the compact set S according to its definition 51. By the Weierstrass theorem, $g(\boldsymbol{\xi}, \boldsymbol{\xi}_t)$ has minimum on S. Using 64, this minimum is global minimum on $\mathbb{R}^d$. Hence $\mathcal{A}(\boldsymbol{\xi}_t)$ is nonempty at any $\boldsymbol{\xi}_t$.

   (b) $h(\mathbf{x}, \mathbf{y})$ is continuous:
   $\mathrm{E}_1(\boldsymbol{\xi})$ is continuous. $\nabla_{\boldsymbol{\xi}} \mathrm{E}_2(\boldsymbol{\xi}) = \boldsymbol{X} \operatorname{softmax}\left(\beta \boldsymbol{X}^T \boldsymbol{\xi}\right)$ is continuous. Hence $g(\mathbf{x}, \mathbf{y})$ is continuous at any $\mathbf{x}, \mathbf{y}$ according to its definition 51. Then $h(\mathbf{x}, \mathbf{y})$ is continuous.

   According to Lemma 12, $\mathcal{A}$ is closed on $\mathbb{R}^d$.

5. **Limit Points are Fixed Points**

   Since

   (a) All points $\boldsymbol{\xi}_{t+1}, t \geqslant 0$ are in the compact set S using 64.

   (b) $\mathcal{A}$ is strictly monotonic with respect to E.

   (c) $\mathcal{A}$ is closed on $\mathbb{R}^d$.

Using Lemma 14, all limit points of the sequence $\{\boldsymbol{\xi}_t\}_{t=0}^{\infty}$ are fixed points of 52.

6. **Limit Points are Stationary Points**

   Using Lemma 15, all fixed points of 52 are stationary points of 50.

7. **Adjacent Point Differences are Convergent**

   Using Lemma 16

$$\left\|\boldsymbol{\xi}_{t+1} - \boldsymbol{\xi}_t\right\|_2 \to 0, \text{ as } t \to \infty \tag{136}$$

8. **Limit Points are Compact**

   Using Lemma 17, the set of limit points of $\{\boldsymbol{\xi}_t\}_{t=0}^{\infty}$ is compact.

9. **Limit Points are Connected**

   Using Lemma 18, if $\{\boldsymbol{\xi}_t\}_{t=0}^{\infty}$ does not converge, then the set of limit points of $\{\boldsymbol{\xi}_t\}_{t=0}^{\infty}$ is connected.

10. **Convergence Sufficient Condition**

    Suppose $\{\boldsymbol{\xi}_t\}_{t=0}^{\infty}$ does not converge, then the set of limit points of $\{\boldsymbol{\xi}_t\}_{t=0}^{\infty}$ must contain at least two limit points, and is connected.

    Since $\mathcal{L}(\mathrm{E}^*)$ is finite, then the set of limit points can be decomposed into the union of two nonempty closed sets of points without common points. Hence it is not connected, which is a contradiction.

    Therefore, if $\mathcal{L}(\mathrm{E}^*)$ is finite, then any sequence $\{\boldsymbol{\xi}_t\}_{t=0}^{\infty}$ generated by the iteration 23 converges to some $\boldsymbol{\xi}^* \in \mathcal{L}(\mathrm{E}^*)$.

□

References

[1] Hubert Ramsauer et al. "Hopfield networks is all you need". In: *arXiv preprint arXiv:2008.02217* (2020).

[2] Alan L Yuille and Anand Rangarajan. "The concave-convex procedure (CCCP)". In: *Advances in neural information processing systems* 14 (2001).

[3] Bharath K Sriperumbudur and Gert RG Lanckriet. "On the Convergence of the Concave-Convex Procedure." In: *Nips*. Vol. 9. Citeseer. 2009, pp. 1759–1767.

[4] Asela Gunawardana, William Byrne, and Michael I Jordan. "Convergence Theorems for Generalized Alternating Minimization Procedures." In: *Journal of machine learning research* 6.12 (2005).

[5] W.I. Zangwill. *Nonlinear Programming: A Unified Approach*. Prentice-Hall international series in management. Prentice-Hall, 1969. ISBN: 9780136235798.

[6] Robert R Meyer. "Sufficient conditions for the convergence of monotonic mathematicalprogramming algorithms". In: *Journal of computer and system sciences* 12.1 (1976), pp. 108–121.