Analysis for the Slow Convergence in Arimoto Algorithm

Kenji Nakagawa
Nagaoka University of Technology
Niigata, Japan 940-2188
Email: nakagawa@nagaokaut.ac.jp

Yoshinori Takei
National Institute of Technology, Akita College
Akita, Japan 011-8511
Email: ytakei@akita-nct.ac.jp

Kohei Watabe
Nagaoka University of Technology
Niigata, Japan 940-2188
Email: k_watabe@vos.nagaokaut.ac.jp

***Abstract*—In this paper, we investigate the convergence speed of the Arimoto algorithm. By analyzing the Taylor expansion of the defining function of the Arimoto algorithm, we will clarify the conditions for the exponential or $1/N$ order convergence and calculate the convergence speed. We show that the convergence speed of the $1/N$ order is evaluated by the derivatives of the Kullback-Leibler divergence with respect to the input probabilities. The analysis for the convergence of the $1/N$ order is new in this paper. Based on the analysis, we will compare the convergence speed of the Arimoto algorithm with the theoretical values obtained in our theorems.**

I. Introduction

Arimoto [2] proposed a sequential algorithm for calculating the channel capacity C of a discrete memoryless channel. Based on the Bayes probability, the algorithm is given by the alternating minimization between the input probabilities and the reverse channel matrices. For arbitrary channel matrix Φ the convergence of the Arimoto algorithm is proved and the convergence speed is evaluated. In the worst case, the convergence speed is the $1/N$ order, and if the input distribution $\boldsymbol{\lambda}^*$ that achieves the channel capacity C is in the interior of the set $\Delta(\mathcal{X})$ of input distributions, the convergence is exponential.

In this paper, we first consider the exponential convergence and evaluate the convergence speed. We show that there exist cases of exponential convergence even if $\boldsymbol{\lambda}^*$ is on the boundary of $\Delta(\mathcal{X})$. Moreover, we also consider the convergence of the $1/N$ order, which is not dealt with in the previous studies. Especially, when the input alphabet size $m=3$, we will analyze the convergence of the $1/N$ order in detail and the convergence speed is evaluated by the derivatives of the Kullback-Leibler divergence with respect to the input probabilities.

As a basic idea for evaluating the convergence speed, we consider that the function F which defines the Arimoto algorithm is a differentiable mapping from $\Delta(\mathcal{X})$ to $\Delta(\mathcal{X})$, and notice that the capacity achieving input distribution $\boldsymbol{\lambda}^*$ is the fixed point of F. Then, the convergence speed is evaluated by analyzing the Taylor expansion of F about the fixed point $\boldsymbol{\lambda}=\boldsymbol{\lambda}^*$.

II. Related works

There have been many related works on the Arimoto algorithm. For example, extension to different types of channels, acceleration of the Arimoto algorithm, characterization of Arimoto algorithm by divergence geometry, etc. If we focus on the analysis for the convergence speed of the Arimoto algorithm, we see in [2],[4],[6] that the eigenvalues of the Jacobian matrix are calculated and the convergence speed is investigated in the case that $\boldsymbol{\lambda}^*$ is in the interior of $\Delta(\mathcal{X})$.

In this paper, we consider the Taylor expansion of the defining function of the Arimoto algorithm. We will calculate not only the Jacobian matrix of the first order term of the Taylor expansion, but also the Hessian matrix of the second order term, and examine the convergence speed of the exponential or $1/N$ order based on the Jacobian and Hessian matrices.

III. Channel matrix and channel capacity

Consider a discrete memoryless channel $X \rightarrow Y$ with the input source X and the output source Y. Let the input alphabet be $\mathcal{X}=\{x_1,\cdots,x_m\}$ and the output alphabet be $\mathcal{Y}=\{y_1,\cdots,y_n\}$.

The conditional probability that the output symbol y_j is received when the input symbol x_i was transmitted is denoted by $P^i_j=P(Y=y_j|X=x_i)$, $i=1,\cdots,m, j=1,\cdots,n$, and the row vector P^i is defined by $P^i=(P^i_1,\cdots,P^i_n)$, $i=1,\cdots,m$. The channel matrix Φ is defined by

$$\Phi=\begin{pmatrix} P^1 \\ \vdots \\ P^m \end{pmatrix}=\begin{pmatrix} P^1_1 & \cdots & P^1_n \\ \vdots & & \vdots \\ P^m_1 & \cdots & P^m_n \end{pmatrix}. \tag{1}$$

The set of input probability distributions on the input alphabet $\mathcal{X}$ is denoted by $\Delta(\mathcal{X})=\{\boldsymbol{\lambda}=(\lambda_1,\cdots,\lambda_m)|\lambda_i \geq 0, i=1,\cdots,m,\sum_{i=1}^m \lambda_i=1\}$. Similarly, the set of output probability distributions on the output alphabet $\mathcal{Y}$ is denoted by $\Delta(\mathcal{Y})=\{Q=(Q_1,\cdots,Q_n)|Q_j\geq 0, j=1,\cdots,n,\sum_{j=1}^n Q_j=1\}$.

Let $Q=\boldsymbol{\lambda}\Phi$ be the output distribution for the input distribution $\boldsymbol{\lambda}\in\Delta(\mathcal{X})$, where the representation by components is $Q_j=\sum_{i=1}^m \lambda_i P^i_j$, $j=1,\cdots,n$, then the mutual information is defined by $I(\boldsymbol{\lambda},\Phi)=\sum_{i=1}^m\sum_{j=1}^n \lambda_i P^i_j \log P^i_j/Q_j$. The channel capacity C is defined by

$$C=\max_{\boldsymbol{\lambda}\in\Delta(\mathcal{X})} I(\boldsymbol{\lambda},\Phi). \tag{2}$$

The Kullback-Leibler divergence $D(Q\|Q')$ for two output distributions $Q=(Q_1,\cdots,Q_n)$, $Q'=(Q'_1,\cdots,Q'_n)\in\Delta(\mathcal{Y})$ is defined by

$$D(Q\|Q')=\sum_{j=1}^n Q_j \log\frac{Q_j}{Q'_j}. \tag{3}$$

The Kullback-Leibler divergence satisfies $D(Q\|Q')\geq 0$, and $D(Q\|Q')=0$ if and only if $Q=Q'$ [3].

An important proposition for investigating the convergence speed of the Arimoto algorithm is the Kuhn-Tucker condition on the input distribution $\boldsymbol{\lambda} = \boldsymbol{\lambda}^*$ to achieve the maximum of (2).

Theorem (Kuhn-Tucker condition) In the maximization problem (2), a necessary and sufficient condition for the input distribution $\boldsymbol{\lambda}^* = (\lambda_1^*, \cdots, \lambda_m^*) \in \Delta(\mathcal{X})$ to achieve the maximum is that there is a certain constant C_0 with

$$D(P^i \| \boldsymbol{\lambda}^* \Phi) \begin{cases} = C_0, & \text{for } i \text{ with } \lambda_i^* > 0, \\ \leq C_0, & \text{for } i \text{ with } \lambda_i^* = 0. \end{cases} \tag{4}$$

In (4), C_0 is equal to the channel capacity C.

Since this Kuhn-Tucker condition is a necessary and sufficient condition, all the information about the capacity achieving input distribution $\boldsymbol{\lambda}^*$ can be derived from this condition.

IV. Arimoto algorithm for calculating channel capacity

A. Arimoto algorithm [2]

A sequence $\{\boldsymbol{\lambda}^N = (\lambda_1^N, \cdots, \lambda_m^N)\}_{N=0,1,\cdots}$ in $\Delta(\mathcal{X})$ is defined by the Arimoto algorithm as follows. First, let $\boldsymbol{\lambda}^0 = (\lambda_1^0, \cdots, \lambda_m^0)$ be an initial distribution taken in the interior of $\Delta(\mathcal{X})$, i.e., $\lambda_i^0 > 0,\ i = 1, \cdots, m$. Then, the Arimoto algorithm is given by the following recurrence formula;

$$\lambda_i^{N+1} = \frac{\lambda_i^N \exp D(P^i \| \boldsymbol{\lambda}^N \Phi)}{\displaystyle\sum_{k=1}^{m} \lambda_k^N \exp D(P^k \| \boldsymbol{\lambda}^N \Phi)}, \quad \begin{array}{l} i = 1, \cdots, m, \\ N = 0, 1, \cdots. \end{array} \tag{5}$$

Let $F_i(\boldsymbol{\lambda})$ be the function on the right hand side of (5). Define $F(\boldsymbol{\lambda}) = (F_1(\boldsymbol{\lambda}), \cdots, F_m(\boldsymbol{\lambda}))$, then we can consider that F is a differentiable mapping from $\Delta(\mathcal{X})$ to $\Delta(\mathcal{X})$, and (5) is represented by

$$\boldsymbol{\lambda}^{N+1} = F(\boldsymbol{\lambda}^N),\ N = 0, 1, \cdots. \tag{6}$$

B. Examples of convergence speed

For many channel matrices Φ, the convergence of the Arimoto algorithm is exponential, but for some special Φ the convergence is very slow. Let us see the situation with the examples below.

First, we classify the indices $i\,(i = 1, \cdots, m)$ in the Kuhn-Tucker condition (4) into 3 types as follows;

$$D(P^i \| \boldsymbol{\lambda}^* \Phi) \begin{cases} = C, & \text{for } i \text{ with } \lambda_i^* > 0, \ \text{(type I)} \\ = C, & \text{for } i \text{ with } \lambda_i^* = 0, \ \text{(type II)} \\ < C, & \text{for } i \text{ with } \lambda_i^* = 0. \ \text{(type III)} \end{cases} \tag{7}$$

Type I index always exists for any channel matrix. On the other hand, type II and type III index may or may not exist.

Now, consider the examples below, taking types I, II, III into account, when the input alphabet size $m = 3$ and the output alphabet size $n = 3$.

1) Example 1 (only type I): If only type I indices exist, then $\lambda_i^* > 0,\ i = 1, 2, 3$, hence $Q^* \equiv \boldsymbol{\lambda}^* \Phi$ is in the interior of $\triangle P^1 P^2 P^3$. We have $D(P^i \| Q^*) = C,\ i = 1, 2, 3$. See Fig.1.

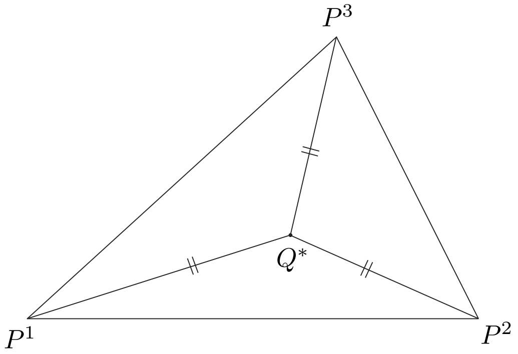

Fig.1. Positional relation of P^1, P^2, P^3 and Q^* of Example 1

As a concrete channel matrix of this example, let us consider

$$\Phi^{(1)} = \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.25 & 0.25 & 0.5 \end{pmatrix}. \tag{8}$$

2) Example 2 (types I and II): If there are type I and type II indices, we can assume $\lambda_1^* > 0, \lambda_2^* > 0, \lambda_3^* = 0$ without loss of generality. We have $D(P^i \| Q^*) = C,\ i = 1, 2, 3$, hence Q^* is just on the side $P^1 P^2$. See Fig.2.

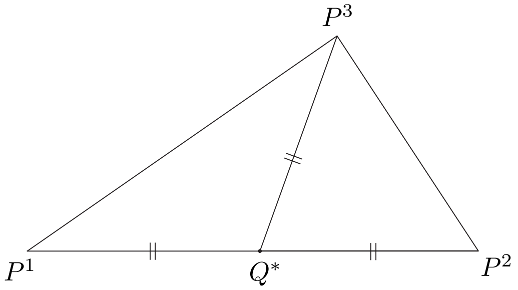

Fig.2. Positional relation of P^1, P^2, P^3 and Q^* of Example 2

As a concrete channel matrix of this example, let us consider

$$\Phi^{(2)} = \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.3 & 0.3 & 0.4 \end{pmatrix}. \tag{9}$$

3) Example 3 (types I and III): If there are type I and type III indices, we can assume $\lambda_1^* > 0, \lambda_2^* > 0, \lambda_3^* = 0$ without loss of generality. We have $C = D(P^1 \| Q^*) = D(P^2 \| Q^*) > D(P^3 \| Q^*)$. See Fig.3.

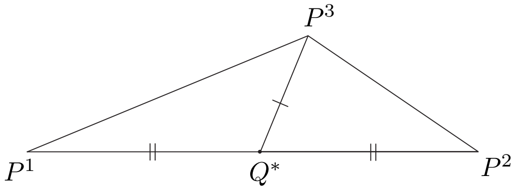

Fig.3. Positional relation of P^1, P^2, P^3 and Q^* of Example 3

As a concrete channel matrix of this example, let us consider

$$\Phi^{(3)} = \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.35 & 0.35 & 0.3 \end{pmatrix}. \tag{10}$$

For the above $\Phi^{(1)}, \Phi^{(2)}, \Phi^{(3)}$, Fig.4 shows the state of convergence of $|\lambda_1^N - \lambda_1^*| \to 0$. By this Figure, we see that in Examples 1 and 3 the convergence is exponential, while in Example 2 the convergence is slower than exponential.

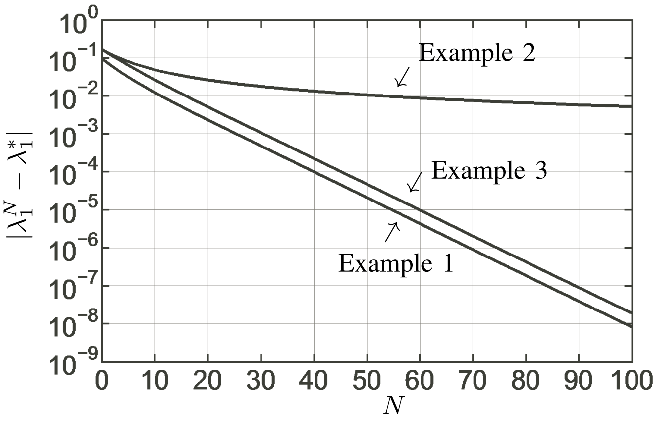

Fig.4. Comparison of the convergence speed of Examples 1,2,3

From the above three examples, it is inferred that the Arimoto algorithm converges very slowly when type II index exists, and converges exponentially when type II index does not exist. We will analyze this phenomenon in the following.

V. Analysis of convergence speed

In this paper, for the analysis of convergence speed, we assume $\operatorname{rank}\Phi = m$.

Lemma 1: Under the assumption of $\operatorname{rank}\Phi = m$, the capacity achieving input distribution $\boldsymbol{\lambda}^*$ is unique and is the fixed point of the mapping F in (6). That is, $\boldsymbol{\lambda}^* = F(\boldsymbol{\lambda}^*)$.

VI. Taylor expansion of $F(\boldsymbol{\lambda})$ about $\boldsymbol{\lambda} = \boldsymbol{\lambda}^*$

The sequence $\boldsymbol{\lambda}^N$ of the Arimoto algorithm converges to the fixed point $\boldsymbol{\lambda}^*$, thus we will examine the convergence speed by the Taylor expansion about the fixed point $\boldsymbol{\lambda} = \boldsymbol{\lambda}^*$. Taylor expansion of the function $F(\boldsymbol{\lambda}) = (F_1(\boldsymbol{\lambda}), \cdots, F_m(\boldsymbol{\lambda}))$ about $\boldsymbol{\lambda} = \boldsymbol{\lambda}^*$ is

$$\begin{aligned} F(\boldsymbol{\lambda}) &= F(\boldsymbol{\lambda}^*) + (\boldsymbol{\lambda} - \boldsymbol{\lambda}^*)J(\boldsymbol{\lambda}^*) \\ &+ \frac{1}{2!}(\boldsymbol{\lambda} - \boldsymbol{\lambda}^*)H(\boldsymbol{\lambda}^*)\,{}^t(\boldsymbol{\lambda} - \boldsymbol{\lambda}^*) + o(\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^*\|^2), \end{aligned} \tag{11}$$

where ${}^t\boldsymbol{\lambda}$ represents the transpose of $\boldsymbol{\lambda}$ and $\|\boldsymbol{\lambda}\|$ represents the Euclidean norm $\|\boldsymbol{\lambda}\| = \left(\lambda_1^2 + \cdots + \lambda_m^2\right)^{1/2}$. $J(\boldsymbol{\lambda}^*)$ is the Jacobian matrix at $\boldsymbol{\lambda} = \boldsymbol{\lambda}^*$, i.e.,

$$J(\boldsymbol{\lambda}^*) = \left(\left.\frac{\partial F_i}{\partial \lambda_{i'}}\right|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^*} \right)_{i',i=1,\cdots,m}, \tag{12}$$

and $H(\boldsymbol{\lambda}^*) = (H_1(\boldsymbol{\lambda}^*), \cdots, H_m(\boldsymbol{\lambda}^*))$, where $H_i(\boldsymbol{\lambda}^*)$ is the Hessian matrix of F_i at $\boldsymbol{\lambda} = \boldsymbol{\lambda}^*$, i.e.,

$$H_i(\boldsymbol{\lambda}^*) = \left(\left.\frac{\partial^2 F_i}{\partial \lambda_{i'} \partial \lambda_{i''}}\right|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^*} \right)_{i',i''=1,\cdots,m}, \tag{13}$$

In (11), $(\boldsymbol{\lambda} - \boldsymbol{\lambda}^*)H(\boldsymbol{\lambda}^*)\,{}^t(\boldsymbol{\lambda} - \boldsymbol{\lambda}^*)$ is an abbreviated expression of the m dimensional vector $((\boldsymbol{\lambda} - \boldsymbol{\lambda}^*)H_1(\boldsymbol{\lambda}^*)\,{}^t(\boldsymbol{\lambda} - \boldsymbol{\lambda}^*), \cdots, (\boldsymbol{\lambda} - \boldsymbol{\lambda}^*)H_m(\boldsymbol{\lambda}^*)\,{}^t(\boldsymbol{\lambda} - \boldsymbol{\lambda}^*))$.

Substituting $\boldsymbol{\lambda} = \boldsymbol{\lambda}^N$ into (11), then by $F(\boldsymbol{\lambda}^*) = \boldsymbol{\lambda}^*$ and $F(\boldsymbol{\lambda}^N) = \boldsymbol{\lambda}^{N+1}$, we have

$$\begin{aligned} &\boldsymbol{\lambda}^{N+1} = \boldsymbol{\lambda}^* + (\boldsymbol{\lambda}^N - \boldsymbol{\lambda}^*)J(\boldsymbol{\lambda}^*) \\ &+ \frac{1}{2!}(\boldsymbol{\lambda}^N - \boldsymbol{\lambda}^*)H(\boldsymbol{\lambda}^*)\,{}^t(\boldsymbol{\lambda}^N - \boldsymbol{\lambda}^*) + o(\|\boldsymbol{\lambda}^N - \boldsymbol{\lambda}^*\|^2). \end{aligned} \tag{14}$$

Then, by putting $\boldsymbol{\mu}^N \equiv \boldsymbol{\lambda}^N - \boldsymbol{\lambda}^*$, (14) becomes

$$\boldsymbol{\mu}^{N+1} = \boldsymbol{\mu}^N J(\boldsymbol{\lambda}^*) + \frac{1}{2!}\boldsymbol{\mu}^N H(\boldsymbol{\lambda}^*)\,{}^t\boldsymbol{\mu}^N + o\left(\|\boldsymbol{\mu}^N\|^2\right). \tag{15}$$

We will investigate the convergence speed of $\boldsymbol{\mu}^N \to \mathbf{0}$.

Writing $\boldsymbol{\mu}^N = (\mu_1^N, \cdots, \mu_m^N)$ with $\mu_i^N \equiv \lambda_i^N - \lambda_i^*$, $i = 1, \cdots, m$, we have $\sum_{i=1}^m \mu_i^N = 0$.

VII. On Jacobian matrix $J(\boldsymbol{\lambda}^*)$

Let us consider the Jacobian matrix $J(\boldsymbol{\lambda}^*)$ for arbitrary m, n. First, we will calculate the components (12) of $J(\boldsymbol{\lambda}^*)$.

Define $D_i \equiv D(P^i \| \boldsymbol{\lambda}\Phi)$, $i = 1, \cdots, m$, for $\boldsymbol{\lambda} \in \Delta(\mathcal{X})$. Further, $Q^* \equiv \boldsymbol{\lambda}^*\Phi$, $D_i^* \equiv D(P^i \| Q^*)$, $i = 1, \cdots, m$, and $D_{i',i}^* \equiv \partial D_i / \partial \lambda_{i'}|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^*}$, $i', i = 1, \cdots, m$. In the Kuhn-Tucker condition (4), define m_1 as the number of indices i with $\lambda_i^* > 0$, i.e., $\lambda_i^* > 0$ for $i = 1, \cdots, m_1$, and $\lambda_i^* = 0$ for $i = m_1 + 1, \cdots, m$.

Then we have

Theorem 1:

$$\begin{aligned} \left.\frac{\partial F_i}{\partial \lambda_{i'}}\right|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^*} &= e^{D_i^* - C}\left(\delta_{i'i} + \lambda_i^* D_{i',i}^*\right) + \lambda_i^*\left(1 - e^{D_{i'}^* - C}\right), \\ &\quad i', i = 1, \cdots, m, \\ &= \begin{cases} \delta_{i'i} + \lambda_i^*\left(D_{i',i}^* + 1 - e^{D_{i'}^* - C}\right), \\ \quad i' = 1, \cdots, m,\ i = 1, \cdots, m_1, \\ e^{D_i^* - C}\delta_{i'i}, \\ \quad i' = 1, \cdots, m,\ i = m_1 + 1, \cdots, m, \end{cases} \end{aligned} \tag{16}$$

where $\delta_{i'i}$ is the Kronecker delta.

Especially, if $D_i^* = C$ for all $1 \le i \le m$, then

$$\left.\frac{\partial F_i}{\partial \lambda_{i'}}\right|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^*} = \begin{cases} \delta_{i'i} + \lambda_i^* D_{i',i}^*, \\ \quad i' = 1, \cdots, m,\ i = 1, \cdots, m_1, \\ \delta_{i'i}, \\ \quad i' = 1, \cdots, m,\ i = m_1 + 1, \cdots, m. \end{cases} \tag{17}$$

A. Eigenvalues of Jacobian matrix $J(\boldsymbol{\lambda}^*)$

From (16), $\partial F_i / \partial \lambda_{i'}|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^*} = 0$ for $i' = 1, \cdots, m_1$, $i = m_1 + 1, \cdots, m$, thus, the Jacobian matrix $J(\boldsymbol{\lambda}^*)$ is of the form

$$\begin{aligned} J(\boldsymbol{\lambda}^*) &\equiv \begin{pmatrix} J^1 & O \\ A & J^2 \end{pmatrix}, \\ &J^1 \in \mathbb{R}^{m_1 \times m_1},\ J^2 \in \mathbb{R}^{(m-m_1)\times(m-m_1)}, \\ &O \in \mathbb{R}^{m_1 \times (m-m_1)},\ A \in \mathbb{R}^{(m-m_1)\times m_1}, \end{aligned} \tag{18}$$

where O denotes the zero matrix. Let $\theta_1, \cdots, \theta_m$ be the eigenvalues of $J(\boldsymbol{\lambda}^*)$. By (18), we put $\theta_1, \cdots, \theta_{m_1}$ to be the eigenvalues of J^1, and $\theta_{m_1+1}, \cdots, \theta_m$ the eigenvalues of J^2.

1) Eigenvalues of J^1: Let $I \in \mathbb{R}^{m_1 \times m_1}$ denote the identity matrix and let $B \equiv I - J^1$. Writing the eigenvalues of B as $\beta_1 \geq \cdots \geq \beta_{m_1}$, then the eigenvalues of J^1 are $\theta_i = 1-\beta_i$, $i = 1, \cdots, m_1$. Define $\Lambda \equiv \mathrm{diag}(\lambda_1^*, \cdots, \lambda_{m_1}^*)$, which is the diagonal matrix with diagonal components $\lambda_1^*, \cdots, \lambda_{m_1}^*$.

Lemma 2: $\sqrt{\Lambda}B\sqrt{\Lambda}^{-1}$ is symmetric and positive definite. Every component of B is non-negative. Every row sum of B is 1. (See [6].)

From Lemma 2 and the Perron-Frobenius theorem, we have

Theorem 2: The eigenvalues of J^1 satisfy

$$0 = \theta_1 \leq \theta_2 \leq \cdots \leq \theta_{m_1} < 1. \tag{19}$$

2) Eigenvalues of J^2: From (16), (18), we have

$$J^2 = \mathrm{diag}\left(e^{D_{m_1+1}^* - C}, \cdots, e^{D_m^* - C}\right) \in \mathbb{R}^{(m-m_1)\times(m-m_1)}.$$

Theorem 3: The eigenvalues of J^2 are $\theta_i = e^{D_i^* - C}$, $i = m_1 + 1, \cdots, m$, and we have

$$0 < \theta_i \leq 1, \, i = m_1 + 1, \cdots, m. \tag{20}$$

VIII. On convergence speed

We obtained in Theorems 2 and 3 the evaluation for the eigenvalues of J^1 and J^2. Let $\theta_{\max} \equiv \max_{1 \leq i \leq m} \theta_i$ be the maximum eigenvalue of $J(\boldsymbol{\lambda}^*)$, then $0 < \theta_{\max} \leq 1$ by Theorems 2 and 3. We separate the case into $0 < \theta_{\max} < 1$ and $\theta_{\max} = 1$. In the following, we will see that $\boldsymbol{\lambda}^N \to \boldsymbol{\lambda}^*$ or $\boldsymbol{\mu}^N \to \mathbf{0}$ is exponential if $0 < \theta_{\max} < 1$, and $1/N$ order convergence if $\theta_{\max} = 1$.

A. Convergence speed in case of $0 < \theta_{\max} < 1$

Theorem 4: Suppose that the maximum eigenvalue $\theta_{\max}$ of the Jacobian matrix $J(\boldsymbol{\lambda}^*)$ satisfies $0 < \theta_{\max} < 1$. Then, for any θ with $\theta_{\max} < \theta < 1$, there exist $\delta > 0$ and $K > 0$, such that for arbitrary initial vector $\boldsymbol{\lambda}^0$ with $\|\boldsymbol{\lambda}^0 - \boldsymbol{\lambda}^*\| < \delta$, we have

$$\|\boldsymbol{\mu}^N\| = \|\boldsymbol{\lambda}^N - \boldsymbol{\lambda}^*\| < K\theta^N, \, N = 0, 1, \cdots, \tag{21}$$

where θ^N denotes the Nth power of θ.

B. Convergence speed in case of $\theta_{\max} = 1$

Because all the eigenvalues of J^1 are smaller than 1 by Theorem 2, $\theta_{\max} = 1$ implies that some eigenvalue of J^2 is 1. That is, $D(P^i \| Q^*) = C$ holds for some i with $m_1 + 1 \leq i \leq m$. Therefore, in this case there exists type II index in (7). In other words, if type II index does not exist the convergence is exponential. Theorem 4 cannot be applied for this case of $\theta_{\max} = 1$. The convergence $\boldsymbol{\mu}^N \to \mathbf{0}$ is not determined only by the Jacobian matrix, but it is necessary to investigate the Hessian matrix in the second order term of the Taylor expansion.

IX. On Hessian matrix

In the previous studies, say, [2],[4],[6], the Jacobian matrix is considered but the Hessian matrix is not. Analysis by the Hessian matrix is new in this paper. Let us calculate the components (13) of the Hessian matrix of the function F_i, $i = 1, \cdots, m$ at $\boldsymbol{\lambda} = \boldsymbol{\lambda}^*$.

Define $D_{i,i',i''}^* \equiv \partial^2 D_i / \partial\lambda_{i'} \partial\lambda_{i''}|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^*}$. We show only the result of calculation.

Theorem 5:

$$\begin{gathered}\left.\frac{\partial^2 F_i}{\partial\lambda_{i'}\partial\lambda_{i''}}\right|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^*} = e^{D_i^* - C}[(1 - e^{D_{i'}^* - C} + D_{i,i'}^*)(\delta_{ii''} + \\ \lambda_i^*(1 - e^{D_{i''}^* - C})) + (1 - e^{D_{i''}^* - C} + D_{i,i''}^*)(\delta_{ii'} + \lambda_i^*(1 - \\ e^{D_{i'}^* - C})) + \lambda_i^*(D_{i,i'}^* D_{i,i''}^* + D_{i,i',i''}^* + D_{i',i''}^* - e^{D_{i'}^* - C} D_{i',i''}^* - \\ e^{D_{i''}^* - C} D_{i',i''}^* - \textstyle\sum_{k=1}^{m_1} \lambda_k^* D_{k,i'}^* D_{k,i''}^*)], i, i', i'' = 1, \cdots, m.\end{gathered}$$

Especially, if $D_i^* = C$ for all $1 \leq i \leq m$,

$$\begin{gathered}\left.\frac{\partial^2 F_i}{\partial\lambda_{i'}\partial\lambda_{i''}}\right|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^*} = \delta_{ii'} D_{i,i''}^* + \delta_{ii''} D_{i,i'}^* + \lambda_i^*(D_{i,i'}^* D_{i,i''}^* + \\ D_{i,i',i''}^* - D_{i',i''}^* - \textstyle\sum_{k=1}^{m_1} \lambda_k^* D_{k,i'}^* D_{k,i''}^*), \, i, i', i'' = 1, \cdots, m,\end{gathered}$$

hence, in this case for $i = m_1 + 1, \cdots, m$,

$$\left.\frac{\partial^2 F_i}{\partial\lambda_{i'}\partial\lambda_{i''}}\right|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^*} = \delta_{ii'} D_{i,i''}^* + \delta_{ii''} D_{i,i'}^*, \; i', i'' = 1, \cdots, m,$$

which is a relatively simple form.

X. Convergence speed in case of $m = 3$ and n is arbitrary

In Theorem 5, the Hessian matrix is very complicated, thus it is difficult to investigate arbitrary channel matrix. Therefore, in this section, we will consider a special case, i.e., $m = 3$ and n is arbitrary. For $m = 3$, without loss of generality, we have the following exhaustive classification.

(i) $\lambda_1^* > 0, \lambda_2^* > 0, \lambda_3^* > 0$,
(ii) $\lambda_1^* > 0, \lambda_2^* > 0, \lambda_3^* = 0, D_3^* = C$,
(iii) $\lambda_1^* > 0, \lambda_2^* > 0, \lambda_3^* = 0, D_3^* < C$.

(i) is the case of Example 1. We have $m = m_1 = 3$ and $J(\boldsymbol{\lambda}^*) = J^1$. Then $0 < \theta_{\max} < 1$, hence by Theorem 4 the convergence $\boldsymbol{\mu}^N \to \mathbf{0}$ is exponential.

In (ii), (iii), we have $m_1 = 2$ and hence

$$\begin{gathered}J(\boldsymbol{\lambda}^*) = \begin{pmatrix} J^1 & O \\ A & J^2 \end{pmatrix}, \\ J^1 \in \mathbb{R}^{2\times 2}, \; J^2 = e^{D_3^* - C} \in \mathbb{R}, \\ O \in \mathbb{R}^{2\times 1}, \; A \in \mathbb{R}^{1\times 2}.\end{gathered} \tag{22}$$

Then in (iii), which is the case of Example 3, we have $J^2 < 1$ and $0 < \theta_{\max} < 1$. Therefore, the convergence $\boldsymbol{\mu}^N \to \mathbf{0}$ is exponential by Theorem 4.

So, the rest is (ii), which is the case of Example 2.

A. Convergence of $1/N$ order

In (ii) above, we have $\theta_{\max} = 1$ hence for the analysis of the convergence speed, we will investigate the Hessian matrix in the second order term of the Taylor expansion. By Theorems 1 and 5, we have $J(\boldsymbol{\lambda}^*)$ and $H_3(\boldsymbol{\lambda}^*)$ as

$$J(\boldsymbol{\lambda}^*) = \begin{pmatrix} 1 + \lambda_1^* D_{1,1}^* & \lambda_2^* D_{1,2}^* & 0 \\ \lambda_1^* D_{1,2}^* & 1 + \lambda_2^* D_{2,2}^* & 0 \\ \lambda_1^* D_{1,3}^* & \lambda_2^* D_{2,3}^* & 1 \end{pmatrix}, \tag{23}$$

$$H_3(\boldsymbol{\lambda}^*) = \begin{pmatrix} 0 & 0 & D_{1,3}^* \\ 0 & 0 & D_{2,3}^* \\ D_{1,3}^* & D_{2,3}^* & 2D_{3,3}^* \end{pmatrix}. \tag{24}$$

$H_1(\boldsymbol{\lambda}^*)$ and $H_2(\boldsymbol{\lambda}^*)$ do not affect directly on the convergence speed.

By the first order term $\boldsymbol{\mu}^{N+1} = \boldsymbol{\mu}^N J(\boldsymbol{\lambda}^*)$ of the Taylor expansion (15) and $\mu_1^N + \mu_2^N + \mu_3^N = 0$, defining

$$\hat{\boldsymbol{\mu}}^N \equiv (\mu_1^N, \mu_2^N),$$
$$\hat{J}(\boldsymbol{\lambda}^*) \equiv \begin{pmatrix} 1+\lambda_1^* D_{1,1}^* - \lambda_1^* D_{1,3}^* & \lambda_2^* D_{1,2}^* - \lambda_2^* D_{2,3}^* \\ \lambda_1^* D_{1,2}^* - \lambda_1^* D_{1,3}^* & 1+\lambda_2^* D_{2,2}^* - \lambda_2^* D_{2,3}^* \end{pmatrix},$$

we have

$$\hat{\boldsymbol{\mu}}^{N+1} = \hat{\boldsymbol{\mu}}^N \hat{J}(\boldsymbol{\lambda}^*). \tag{25}$$

Lemma 3: The eigenvalues of $\hat{J}(\boldsymbol{\lambda}^*)$ are $\theta_1 \equiv -D_{1,2}^*$ and $\theta_2 \equiv 1$. We have $0 \le \theta_1 < 1$.

A right eigenvector of $\hat{J}(\boldsymbol{\lambda}^*)$ for $\theta_1 = -D_{1,2}^*$ is $\boldsymbol{a} = \begin{pmatrix} \lambda_2^*(D_{1,2}^* - D_{2,3}^*) \\ -\lambda_1^*(D_{1,2}^* - D_{1,3}^*) \end{pmatrix}$. Multiplying the both sides of (25) by $\boldsymbol{a}$ from the right, we have $\hat{\boldsymbol{\mu}}^N \boldsymbol{a} = K\left(\theta_1\right)^N$, $K \equiv \hat{\boldsymbol{\mu}}^0 \boldsymbol{a}$ (constant). By solving $\hat{\boldsymbol{\mu}}^N \boldsymbol{a} = K\left(\theta_1\right)^N$ and $\mu_1^N + \mu_2^N = -\mu_3^N$, we have

$$\mu_1^N = -b_1\mu_3^N + K_1(\theta_1)^N,\ \mu_2^N = -b_2\mu_3^N - K_1(\theta_1)^N, \tag{26}$$

where K_1 is a constant and

$$b_1 \equiv \frac{\lambda_1^*(D_{1,2}^* - D_{1,3}^*)}{1+D_{1,2}^*},\ b_2 \equiv \frac{\lambda_2^*(D_{1,2}^* - D_{2,3}^*)}{1+D_{1,2}^*},\ b_1 + b_2 = 1.$$

Now, we consider the third component of (15);

$$\begin{aligned} \mu_3^{N+1} &= \boldsymbol{\mu}^N \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \\ &+ \frac{1}{2!}\left(\mu_1^N, \mu_2^N, \mu_3^N\right) \begin{pmatrix} 0 & 0 & D_{1,3}^* \\ 0 & 0 & D_{2,3}^* \\ D_{1,3}^* & D_{2,3}^* & 2D_{3,3}^* \end{pmatrix} \begin{pmatrix} \mu_1^N \\ \mu_2^N \\ \mu_3^N \end{pmatrix} \\ &+ o\left(\|\boldsymbol{\mu}^N\|^2\right) \\ &= \mu_3^N - \rho\left(\mu_3^N\right)^2 + o\left(|\mu_3^N|^2\right), \end{aligned} \tag{27}$$

where $\rho \equiv D_{1,3}^* b_1 + D_{2,3}^* b_2 - D_{3,3}^*$.

Here, we assume $\rho > 0$. If $\rho < 0$, the recurrence formula (27) diverges, thus we know $\rho \ge 0$. Hence, our assumption implies $\rho \ne 0$.

From (26), (27) and Lemma 3, we have

Theorem 6:

$$\begin{cases} \lim\limits_{N\to\infty} N\mu_1^N = -\dfrac{b_1}{\rho}, & (28) \\ \lim\limits_{N\to\infty} N\mu_2^N = -\dfrac{b_2}{\rho}, & (29) \\ \lim\limits_{N\to\infty} N\mu_3^N = \dfrac{1}{\rho}. & (30) \end{cases}$$

XI. Example of convergence of $1/N$ order

The following example gives the convergence speed of the $1/N$ order.

Example 4: Consider the channel matrix

$$\Phi = \begin{pmatrix} 0.720 & 0.215 & 0.065 \\ 0.013 & 0.431 & 0.556 \\ 0.250 & 0.700 & 0.050 \end{pmatrix},$$

then we have $\boldsymbol{\lambda}^* = (0.453, 0.547, 0.000)$, $Q^* = (0.333, 0.333, 0.334)$,

$$J(\boldsymbol{\lambda}^*) = \begin{pmatrix} 0.227 & -0.227 & 0 \\ -0.188 & 0.188 & 0 \\ -0.453 & -0.547 & 1 \end{pmatrix},$$
$$H_3(\boldsymbol{\lambda}^*) = \begin{pmatrix} 0 & 0 & -1.000 \\ 0 & 0 & -1.000 \\ -1.000 & -1.000 & -3.330 \end{pmatrix}.$$

We have by Theorem 6,

$$\lim_{N\to\infty} N\boldsymbol{\mu}^N = (-0.682, -0.822, 1.504). \tag{31}$$

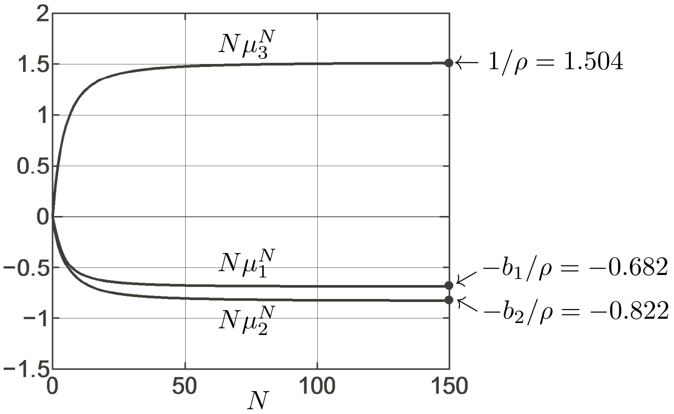

Fig.5. Convergence of $N\mu_i^N$ of Example 4

We see from Fig.5 that (31) holds, i.e., $N\mu_i^N$ converges to the theoretical value obtained in Theorem 6.

XII. Conclusion

In this paper, we investigated the convergence speed of the Arimoto algorithm, especially slow convergence of $1/N$ order by analyzing the Taylor expansion of the defining function. We confirmed that the obtained theoretical values of the convergence speed agreed with the simulation values.

XIII. Acknowledgement

This work was supported by JSPS KAKENHI Grant Number JP17K00008.

References

[1] S. Amari, *Differential-Geometrical Methods in Statistics,* Lecture Notes in Statistics, Springer-Verlag, 1985.

[2] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," IEEE Trans. Inf. Theory, vol. IT-18, pp.14-20, 1972.

[3] I. Csiszàr and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, Orlando, 1982.

[4] G. Matz and P. Duhamel, "Information Geometric Formulation and Interpretation of Accelerated Blahut-Arimoto-Type Algorithms," in Proceedings of ITW2004, 2004.

[5] K. Nakagawa, K. Watabe and T. Sabu, "On the Search Algorithm for the Output Distribution that Achieves the Channel Capacity," IEEE Trans. Inf. Theory, Vol. 63, No. 2, pp.1043-1062, February 2017.

[6] Y. Yu, "Squeezing the Arimoto-Blahut Algorithm for Faster Convergence," IEEE Trans. Inf. Theory, Vol.56, No.7, pp.3149-3157, January 2010.