GENERALIZATION OF THE MEAN-FIELD METHOD FOR POWER-LAW DISTRIBUTIONS

TARO TOYOIZUMI
Department of Complexity Science and Engineering,
Graduate School of Frontier Sciences,
The University of Tokyo, 4-6-1 Komaba,
Meguro-ku, 153-8905 Tokyo, Japan
taro@sat.t.u-tokyo.ac.jp

KAZUYUKI AIHARA
Department of Information and Systems,
Institute of Industrial Science,
The University of Tokyo, 4-6-1 Komaba,
Meguro-ku, 153-8505 Tokyo, Japan
ERATO Aihara Complexity Modelling Project,
JST, 3-23-5 Vehara, Shibuya-ku, 151-0064 Tokyo, Japan
aihara@sat.t.u-tokyo.ac.jp

Received October 31, 2004

Recently much attention has been paid to the nonextensive canonical distributions: the $\alpha$-families. Such distributions have been found in many real-world systems such as fully developed turbulence and financial markets. In this paper, a generalized mean-field method to approximate the expectations of the $\alpha$-families is proposed. We calculate the $\alpha'$-projection of a probability distribution to find that the computational complexity to approximate the expectations is greatly reduced with a proper choice of the projection-index $\alpha'$. We apply this method to a simple binary-state system and compare the results with direct numerical calculations.

Keywords: Information geometry; generalized statistical mechanics; $\alpha$-divergence.

1. Introduction

During the last decade, Generalized Statistical Mechanics (GSM) has been intensively studied [Abe & Okamoto, 2001]. Adding one parameter, Tsallis [1988] proposed a generalized version of Shannon entropy, $S_q = -k[1 - \sum_x p(x)]/(1 - q)$, where $q$ is the entropic index and $p(x)$ is the microscopic probability of a state $x$. As the limit of $q \to 1$, the ordinary Shannon entropy is derived. Maximization of this entropy under an energy constraint yields the GSM canonical distribution [Tsallis, 1988], i.e. $p(x) = (1/Z)(1 - (1 - q)\beta H(x))^{1/(1 - q)}$. GSM describes a large number of important real-world phenomena with this single-parameter generalization such as self-gravitating systems [Taruya & Sakagami, 2002], long-range classical Hamiltonian systems [Latora et al., 1998; Latora et al., 2001; Latora & Tsallis, 2001], fully developed turbulence [Beck, 2001, 2002], financial markets [Ghashghaie et al., 1996], and one-dimensional nonlinear maps [Baldovin & Rovledo, 2002] (see [Abe & Okamoto, 2001] for a summary). The more the importance of

---

1 This paper is a modification of the contribution in the proceedings of Shanghai International Symposium on Nonlinear Science and Applications 2003 (SNSA 2003).
GSM is recognized, the greater the need for tools to analyze the GSM canonical distribution. Because of the correlations among the variables x of the GSM canonical distribution, it is computationally hard to elucidate statistical properties of a large-size system. In this respect some methods, for example, the mean-field method and the variational method have been arranged for GSM [Plastino & Tsallis, 1993; Lenz et al., 1998; Mendes et al., 1999].

On the other hand, the mean-field method is now not only of interest to physicists but also used in the fields of information theory [Kabashima & Saad, 1998] and machine learning [Peterson & Anderson, 1987; Opper & Winther, 2000]. That is why the mean-field method is intensively studied in the framework of the information geometry [Tanaka, 2000; Bhattacharyya & Keerthi, 2000]. Amari et al. proposed the α-projection of the Boltzmann–Gibbs distribution as a generalization of the mean-field method [Amari et al., 2001]. In this paper, however, we apply this method to power-law distributions that are known as α-families in the field of information geometry [Amari & Nagaoka, 2000].

We show that the particular selection of a projection enables us to approximate the expectations of a distribution with less computational complexity compared with the exhaustive exact calculation.

The main purpose of this paper is to compute the expectation η of an α-family:

\[
p(x; \theta) = \begin{cases} 
\frac{1}{Z(\theta)} \exp \left( \sum_{\nu=1}^{N} \theta^\nu f_\nu(x) \right), & \alpha = 1 \\
\frac{1}{Z(\theta)} \left[ 1 - \alpha \sum_{\nu=0}^{N} \theta^\nu f_\nu(x) \right]^{2/(1-\alpha)}, & \alpha \neq 1 
\end{cases}
\]

where \( \{f_\nu\}_{\nu=0}^{N} \) is a set of linear independent functions of the state vector x = \( \{x_1, \ldots, x_N\} \), \( \theta = \{\theta^\nu\}_{\nu=0}^{N} \) is a coordinate system of the α-family, and Z(θ) is the normalization constant.

We assume \( f_0 = 1 \), \( \theta^0 = 2/(1-\alpha) \), and [(1 - \alpha)/2] \( \sum_{\nu=0}^{N} \theta^\nu f_\nu(x) > 0 \) for all x throughout this paper for simplicity. The expectations of p(x; θ) are given as

\[
\eta_\nu \equiv E[p_i] = \sum_{x} f_\nu(x) p(x; \theta),
\]

for \( \nu = 1, \ldots, N \). When \( \sum_{\nu=1}^{N} \theta^\nu f_\nu(x) = -\beta H(x) \) and \( (1 - \alpha)/2 = (1 - q) \), p(x; θ) of Eq. (1) is the GSM canonical distribution.

Due to the nonlinear terms of \( \{f_\nu\} \) in Eq. (1) for \( \alpha \neq 1 \), it is computationally hard to calculate η for large N systems. It takes \( O(Nk^N) \) of computation for general α, where k is the number of discrete states that \( x_i \) can take. Therefore, we usually apply the mean-field method to approximate η. In the following, we propose a generalization of the mean-field method, that is the α-projection [Amari & Nagaoka, 2000; Amari et al., 2001] of the α-family.

2. The α-Projections of Power-Law Distributions

Let x = \( \{x_i | x_i \in \{1, \ldots, k\}, i = 1, \ldots, N\} \) be a state vector, S be the family of probability distributions on x, and \( \mathcal{M} \) be the family of factorizable probability distributions on x. A probability distribution on \( \mathcal{M} \) is represented as \( p(x; \theta) = \prod_{i=1}^{N} p_i(x_i; \theta^i) \), where \( \theta^i \) is a coordinate system of \( \mathcal{M} \). In this section, we approximate p(x; θ) (∈ S) of Eq. (1) by its α-projection onto \( \mathcal{M} \).

We will show in the following that a proper selection of the projection-index \( \alpha' \) can considerably reduce the computation of the α'-projection of p onto \( \mathcal{M} \) [Toyozumi & Aihara, 2003]. Then we will approximate the expectation η by \( \eta_0 \equiv E[p(x; \theta)]_{\nu=1}^{N} \). We will discuss the properties of the α'-projection and explain that this method is a one-parameter generalization of the naive mean-field method.

Let us calculate the α'-divergence \( D_{\alpha'} \) between \( p \) and \( p_0 \) and the \( \alpha' \)-projection of \( p \) onto \( \mathcal{M} \). Since it is given by arg \( \min_{p \in \mathcal{M}} D_{\alpha'}(p[p_0]) \) [See Appendix for properties of \( \alpha' \)-projection,], the α'-divergence between \( p \) and \( p_0 \) is expressed as,

\[
D_{\alpha'}(p[p_0]) = \frac{1}{1 - \alpha'} \left[ 1 - \sum_{x} p^{(1-\alpha')/2} p^{(1+\alpha')/2} \right] = \frac{1}{1 - \alpha'} \left[ 1 - e^{-\frac{1}{1-\alpha'} \log \left( \frac{p}{p_0} \right)} \right],
\]

where \( \psi = \log Z \) and

\[
\frac{1 + \alpha'}{2} \frac{G_{\alpha'}}{2} \equiv -\log Z + \frac{2}{\alpha'} - 1 \times \log \sum_{x} p^{(1-\alpha')/2} p_{0}^{(1+\alpha')/2}.
\]


Note that the second term of (4) is Rényi’s G-divergence [Arndt, 2001]. It is easy to check that \( G_{\alpha} \) is a monotonic increasing function of \( D_{\alpha} \) from Eq. (3), therefore \( \arg\min_{p \in \mathcal{M}} D_{\alpha}(p) = \arg\min_{h \in \mathcal{H}} G_{\alpha}(p || p_0) \). We minimize \( G_{\alpha} \) instead of \( D_{\alpha} \) hereafter.

Then what \( \alpha' \) should we choose to approximate \( \eta \) by \( \eta_\alpha \)? If \( \mathcal{M} \) is a 1-parametral submanifold of \( \mathcal{S} \), we have to calculate the \((-1)\)-projection of \( p \) for the exact expectations. However, the calculation of \((-1)\)-projection is as computationally hard as that of direct calculation [Amari et al., 2001]. Thus we should choose a proper \( \alpha' \) taking account of the computational complexity.

Because \( p \) is a distribution of the \( \alpha \)-family, it is represented as \( p^{(1-\alpha)/2} = c_1 \sum_{\nu=0}^{\infty} \theta^\nu f_\nu + c_2 \) with some constants \( c_1 \) and \( c_2 \). Suppose we choose \( \alpha' = \alpha \) here. In this case, \( G_{\alpha} = \left(-\frac{4}{(1-\alpha^2)}\right) \log \left( \frac{(1-\alpha)/2}{\sum_{\nu=0}^{\infty} \theta^\nu (f_\nu)^{\frac{\alpha}{\nu}} \sum_{\nu=0}^{\infty} \nu f_\nu (x_0)^{1+(\alpha/2)\theta\nu} / (x_0,\mathbf{h}) \right) \). If \( f_{\nu}(x) (\nu = 0, 1, \ldots, N) \) are functions of \( \nu \) variables, it takes \( O(N^\nu) \) steps to calculate \( G_\alpha \). In addition, we can also calculate

\[
\frac{\partial G_\alpha}{\partial \theta^i} = -\frac{4}{1-\alpha^2} \nu \sum_{\nu=0}^{\infty} \frac{\partial}{\partial \theta^i} (f_{\nu})^\frac{\alpha}{\nu}
\]

for \( i = 1, \ldots, N \) at the same orders. One can easily find \( \mathbf{h} \) that gives a local minimum of \( G_{\alpha} \), by applying an optimization algorithm.

In this way, this approximation greatly reduces the number of operations for systems with large \( N \), while the exact calculation of \( \eta \) requires \( O(Nk^N) \) operations in general.

### 3. Application to a Binary-State Model

In this section we calculate the \( \alpha \)-projection of a binary-state distribution: the distribution of Eq. (1) with \( \mathbf{x} = [x_i | x_i \in \{+1, -1\}, i = 1, \ldots, N \) and \( \sum_{i=1}^{N} \theta^\nu f_\nu (x) = \sum_{i=1}^{N} \sum_{j=1}^{N} \theta^\nu x_i x_j + \sum_{i=1}^{N} \theta^\nu x_i \). Here we assume that \( \min_{\nu} \left( (1-\alpha)/2 \sum_{\nu=0}^{\infty} \theta^\nu f_\nu \right) = c > 0 \). We approximate \( p \) by \( p_0 \) that minimizes the \( \alpha \)-divergence between them. Because \( \mathbf{x} \) is a binary vector here, \( p_0 \) is generally represented in the following exponential form:

\[
p_0 = \frac{1}{Z(h)} \exp \left( \sum_{i=1}^{N} h_i x_i \right),
\]

Let us introduce the following expectations,

\[
\eta^{(\alpha)}_\nu \equiv \sum_x f_\nu (x) p_0 \left( \frac{1 + \alpha}{2} h \right)
\]

\[
= \left\{ \begin{array}{ll}
\tanh \left( \frac{1 + \alpha}{2} h \right), & f_\nu = x_i x_j \\
\tanh \left( \frac{1 + \alpha}{2} h \right), & f_\nu = x_i
\end{array} \right.
\]

for \( \nu = 1, \ldots, N \). Then we find the gradient of Eq. (5) to be

\[
\frac{\partial G_\alpha}{\partial \theta^\nu} = \frac{2}{1-\alpha} \left[ \nu g_\nu - g^{(\alpha)}_\nu - g^{(\alpha)}_\nu \sum_{\nu=1}^{N} \frac{\theta^\nu g^{(\alpha)}_\nu}{\nu} \right]
\]

with \( g^{(\alpha)}_\nu = \left[ 1 - (\nu g^{(\alpha)}_\nu)^2 \right] \). We employ the gradient descent algorithm to search for a local minimum of \( G_\alpha \) (Algorithm A in Table 1).

As the stationary condition of Algorithm A for \( \alpha \to 1 \), we can derive the usual self-consistent equations of the naive mean-field method:

\[
\eta_i = \tanh \left( \frac{1 + \alpha}{2} h \right)
\]

for \( i = 1, \ldots, N \), which are equations of the naive mean-field method. The reason is that the naive mean-field equation is derived from the saddle point condition of the 1-divergence. Therefore, in this case, the \( \alpha \)-projection is a generalization of the naive mean-field method for \( \alpha \)-families.

#### 3.1. Numerical results

In this section we apply the method introduced in Sec. 3 to a system with small \( N \) and compare

<table>
<thead>
<tr>
<th>Table 1. Algorithm A</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Initialize ( \mathbf{h} ) to small random values</td>
</tr>
<tr>
<td>2. Calculate ( g^{(\alpha)}<em>\nu ) and ( \partial G</em>\alpha/\partial \theta^\nu ).</td>
</tr>
<tr>
<td>3. Update ( \mathbf{h} ) according to</td>
</tr>
<tr>
<td>( h_{\nu} = h_{\nu} - \delta \frac{\partial G_\alpha}{\partial \theta^\nu} )</td>
</tr>
<tr>
<td>where ( \delta ) is a step size</td>
</tr>
<tr>
<td>4. Return to step 2, stop after finite steps ( n ).</td>
</tr>
</tbody>
</table>
the results with direct numerical calculations. We also compare the results with another method that could be applied: the mean-field approximation of the Callen identity [Sarmento, 1995].

Let us first derive the mean-field approximation of the Callen identity here. The single-site Callen identity is

\[ \eta_i = \frac{\mathbb{E}_p[(p(x_i = 1, \eta) - p(x_i = -1, \eta)]}{\mathbb{E}_p[(p(x_i = 1, \eta) + p(x_i = -1, \eta))]}, \]

where \( \eta \equiv \{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_N \} \). Employing the naive mean-field approximation of the above equation, we obtain the self-consistent equations for \( \eta \):

\[ \eta_i \approx \frac{p(x_i = 1, \eta)}{p(x_i = 1, \eta) + p(x_i = -1, \eta)}, \]

for all \( i \). We can compute the approximation of \( \eta \) by calculating Eq. (9) repeatedly (Algorithm B in Table 2).

Note that Algorithm B also has the same stationary condition as the naive mean-field method when \( \alpha \to 1 \). Therefore for \( \alpha \approx 1 \), their differences only come from the optimization algorithms that we apply.

Figure 1 shows the average of normalized error:

\[ \text{Err} \equiv \frac{\sum_{i=1}^{N} |\eta_i - \eta_{0i}|}{\sum_{i=1}^{N} |\eta_i|} \]

and its standard deviation obtained from Algorithms A (cross) and B (square) for \( N = 10 \) and \( c = 1 \). \( \{\theta^1\} \) and \( \{\theta^2\} \) are generated from the Gauss distributions \( N(\mu, 0.5) \) and \( N(0, 0, 1.0) \), respectively. We generate 100 samples here. Since it is generally difficult to find the step size \( \delta \) of Algorithm A, we repeat this algorithm five times over different \( \delta = \{1.0, 6.0, 11.0\} \) and choose the \( \eta_0 \) that minimizes \( D_\alpha \) after \( n^* = 30 \) iterative steps. We carry out just one series (\( n^{**} = 90 \) iterative steps) for Algorithm B. We show in Fig. 1 that Algorithm A gives better results for every \( \alpha \).
that are shown here. If \( \alpha = -1 \), then Algorithm A has a unique minimum at which \( \eta = \eta_0 \) holds [Amari et al., 2001]. Thus Algorithm A finds the exact expectations in this case, with a sufficiently small step size \( \delta \) and many iterations. On the other hand, since the divergence of Eq. (3) may take extremely large values at \( p \approx 0 \) (resp. \( p_0 \approx 0 \)) for \( \alpha > 1 \) (resp. \( \alpha < -1 \)), it is possible that Algorithm A extracts poor results in these cases.

4. The Choice of Projections

What is important for the approximation of Sec. 2 is a proper choice of the projection. We have previously chosen the projection-index \( \alpha' \) that most alleviates the computational cost to approximate \( \eta \). As we will see, there is a tradeoff between the computational complexity and the accuracy of the approximation.

If \( p(x; \theta) \) is a distribution of the 1-family, i.e. \( \alpha = 1 \), the 1-projection is computationally the easiest, while the \((-1)\)-projection yields the exact expectations [Amari et al., 2001]. As it is generally difficult to estimate to what extent the choice of the projection-index affects the approximation of \( \eta \), we take here the simplest model to compare several projections.

Now we consider the \( \alpha \)-family \( |\alpha| < 1 \) of Eq. (1) with \( x_i, x_{\ell} \in \{+1, -1\}, i, \ell = 1, 2 \) and \( H(x) = \theta_0 x_1 x_2 + \theta_0 x_1 + \theta_0 x_2 \), where \( \theta^2 = (2/(1 - \alpha)) H \) and \( \theta_0 = \theta^2 = (2/(1 - \alpha)) H \).

First, a direct calculation gives the expectation as in Eq. (13).

\[
\eta = \frac{[1 + 2H + J^2(1-\alpha)]}{[1 + 2H + J^2(1-\alpha) + 2H(\alpha^2 - 1) + J^2(1-\alpha)^2]} = \frac{[1 + H(\eta_{\text{MF}}) + 1] + J\eta_{\text{MF}} + 2}{[1 + H(\eta_{\text{MF}}) + 1] + J\eta_{\text{MF}} + 2(1-\alpha)}
\]

\[
G_{\alpha, \eta} = -\frac{1}{1 - \alpha^2} \log(c_n + 2H_n + \eta_{\alpha} + J_n \eta_{\alpha}^2) - (1 + \alpha) \log(2 \cosh h) + 2 \log \left( \frac{2 \cosh \frac{1 + \alpha}{2}}{h} \right)
\]

When \( \alpha \approx 1 \), a larger \( n \) provides a better approximation of \( \eta \) (as long as \( M \) is a 1-autoparallel submanifold of \( S \)). Thus in this case there is a tradeoff between the computational complexity and the precision of the approximation. The method of the \( \alpha' \)-projection provides a parameter \( \alpha' \) that controls this tradeoff.

Second, the naive mean-field approximation of Callen identity, i.e. Eq. (9), is calculated to be Eq. (14).

Finally, for an integer \( n \) and \( \alpha' = 1 - (1 - \alpha)n \), we can also calculate the \( G_{\alpha', \eta} \) of Eq. (4) to be \( G_{\alpha', \eta} \). Eq. (15), where \( \eta_{\alpha'} = \text{tanh}((1 + \alpha')/2h) \), \( (c_n, H_n, J_n) = (1, H, J) \), and

\[
c_n = c_{n-1} + 2H_n c_{n-1} + J_n c_{n-1} - 1,
\]

\[
H_n = H(c_{n-1} + J_n c_{n-1} - 1),
\]

\[
J_n = J_{n-1} + c_{n-1} - 2H_n c_{n-1},
\]

for \( n > 2 \); we choose the \( \eta \) that minimizes Eq. (15).

In Fig. 2 we show the expectations derived from Eq. (13) (Direct), from Eq. (14) (MF), and from Eq. (15) with \( n = \{1, 3, 7\} \) for \( \alpha = 0.6 \) and \( H = 0.1 \). Since \( \alpha' = -1 \) in this case, the \( \alpha' \)-projection gives a good approximation when \( n \approx 5 \).

5. Conclusion

We have shown that it is possible to realize a generalization of the mean-field method by calculating the \( \alpha \)-projection of power-law distributions. The number of operations needed to approximate the
expectations is greatly reduced with a proper choice of the projection. We have applied this method to a simple binary-state $\alpha$-family and compared the method with the mean-field approximation of the Calen identity. As a result of numerical calculations, the generalized mean-field method provides less errors for the expectations compared with the other method, especially when it is applied to $\alpha$-families with $\alpha \approx -1$.

Although we have only considered factorizable distributions to approximate the true distribution, it is possible to obtain better approximation by considering a wider class of distributions. Since there are a lot of useful techniques to deal with such structured distributions in the field of information theory, this is an important problem for future study.

As $\alpha$-families are attracting more and more attention in fields such as fully developed turbulence, economics and self-organized criticality, it is important to study the applications of this method to such complex systems.

Acknowledgments
The authors thank Hidetoshi Shimokawa, Kazuyuki Tanaka, and Yoshiyuki Kabashima for beneficial discussions. This work is supported by the Research Fellowships of the Japan Society for the Promotion of Science for Young Scientists and a Grant-in-Aid for JSPS Fellows. This study is also partially supported by the Advanced and Innovative Research program in Life Sciences and a Grant-in-Aid No. 17022017 on Priority Areas of the Ministry of Education, Culture, Sports, Science, and Technology, from the Japanese Government.

References

Appendix
A.1. Information Geometry
In this Appendix, we briefly review the framework of information geometry. Information geometry describes the way to introduce a geometrical
structure into a space of probability distributions once a divergence is given [Amari & Nagaoka, 2000]. Let $S = \{p(x; \theta)\}$ be an $\alpha$-family. The $\alpha$-divergence between two probability distributions $p = p(x; \theta)$ and $p' = p(x; \theta')$ is defined as

$$D_\alpha(p||p') \equiv \frac{4}{1-\alpha^2} \left[ 1 - \sum_x p^{(1-\alpha)/2} p'^{(1+\alpha)/2} \right]$$

(A.1)

for $\alpha \neq \pm 1$, and $D_{\alpha=1}(p||p') \equiv \lim_{\alpha \to 1} D_\alpha(p||p')$ for $\alpha = \pm 1$. Note that $D_{\alpha=1}$ is equivalent to the well-known Kullback-Leibler divergence. The non-negative property of the $\alpha$-divergence, $D_\alpha(p||p') \geq 0$, is directly derived from Jensen’s inequality, where the equality holds if and only if $p = p'$. It is not symmetric except for $\alpha = 0$ and satisfies $D_\alpha(p||p') = D_{-\alpha}(p'||p)$. The $\alpha$-divergence is a distance-like measure representing the difference between two probability distributions.

When $p$ and $p'$ are close enough, i.e. $\theta' = \theta + \Delta \theta$, we can expand the $\alpha$-divergence between these points as $D_\alpha(p(\theta)||p(\theta + \Delta \theta)) = (1/2) \sum_{\nu, \lambda} g_{\alpha,\lambda}(\theta) d\nu d\lambda + o((d\nu)^2)$, where the Fisher metric $g_{\alpha,\lambda}(\theta) \equiv \lim_{\nu, \lambda \to 0} \partial_\nu \partial_\lambda D_\alpha(p||p')$. $\theta$ is the coordinate system of $S$ and $\partial_\nu \equiv (\partial/\partial \nu^\alpha)$ is the natural basis of $\theta$. Note that $[g_{\alpha,\lambda}]$ is a symmetric and positive definite matrix, invariant to the value of $\alpha$ [Amari & Nagaoka, 2000]. We define the inner product of two basis at $\theta$ as $\langle \partial_\nu, \partial_\lambda \rangle \equiv g_{\alpha,\lambda}(\theta)$. Then we can show the following two theorems (see [Amari & Nagaoka, 2000] for proofs):

**Theorem 1.** Let $S$ be an $\alpha$-family, $p$, $q$ and $r$ be three points in $S$, $\gamma_1$ be an $\alpha$-geodesic connecting $p$ and $q$ and $\gamma_2$ be an $(-\alpha)$-geodesic connecting $q$ and $r$ in $S$. If the curves $\gamma_1$ and $\gamma_2$ are orthogonal at the intersection $q$, then the following equality holds: $D_\alpha(p||r) = D_\alpha(p||q) + D_\alpha(q||r) - ((1-\alpha^2)/4) D_\alpha(p||q) D_\alpha(q||r)$.

**Theorem 2.** Let $S$ be an $\alpha$-family, $M$ a submanifold of $S$, and $p$ a point in $S$. A necessary and sufficient condition for a point $q \in M$ to be a stationary point of the function $r \mapsto D_\alpha(p||r)$ restricted on $M$ is that the $\alpha$-geodesic connecting $p$ and $q$ to be orthogonal to $M$ at $q$. We call such a $q$ as an $\alpha$-projection of $p$ onto $M$. 