International Journal of Bifurcation and Chaos, Vol. 16, No. 1 (2006) 129–135

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, 153-8505 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 α-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 α-families is proposed. We calculate the α'-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 α'. We apply this method to a simple binary-state system and compare the results with direct numerical calculations.[1]

Keywords: Information geometry; generalized statistical mechanics; α-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, $S_q = -k[1 - \sum_{\mathbf{x}} p^q(\mathbf{x})]/(1-q)$, where q is the entropic index and $p(\mathbf{x})$ is the microscopic probability of a state $\mathbf{x}$. 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. $p(\mathbf{x}) = (1/Z)[1 - (1-q)\beta\mathcal{H}(\mathbf{x})]^{1/(1-q)}$. 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 $\mathbf{x}$ 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; Lenzi *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 $\boldsymbol{\eta}$ of an α-family:

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

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

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

$$\eta_\nu \equiv E_p[f_\nu] = \sum_{\mathbf{x}} f_\nu(\mathbf{x})p(\mathbf{x};\boldsymbol{\theta}), \tag{2}$$

for $\nu = 1, \ldots, \tilde{N}$. When $\sum_{\nu=1}^{\tilde{N}}\theta^\nu f_\nu(\mathbf{x}) = -\beta\mathcal{H}(\mathbf{x})$ and $(1-\alpha)/2 = (1-q)$, $p(\mathbf{x};\boldsymbol{\theta})$ 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 $\boldsymbol{\eta}$ 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 $\boldsymbol{\eta}$. 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 $\mathbf{x} = \{x_i | x_i \in \{1, \ldots, k\},\ i = 1, \ldots, N\}$ be a state vector, $\mathcal{S}$ be the family of probability distributions on $\mathbf{x}$, and $\mathcal{M}$ be the family of factorizable probability distributions on $\mathbf{x}$. A probability distribution on $\mathcal{M}$ is represented as $p_0(\mathbf{x};\mathbf{h}) = \prod_{i=1}^N p_{0i}(x_i;h^i)$, where $\mathbf{h} = \{h^i\}_{i=1}^N$ is a coordinate system of $\mathcal{M}$. In this section, we approximate $p(\mathbf{x};\boldsymbol{\theta})(\in \mathcal{S})$ of Eq. (1) by its α'-projection onto $\mathcal{M}$.

We will show in the following that a proper selection of the projection-index α' can considerably reduce the computation of the α'-projection of p onto $\mathcal{M}$ [Toyoizumi & Aihara, 2003]. Then we will approximate the expectation $\boldsymbol{\eta}$ by $\boldsymbol{\eta}_0 \equiv \{E_{p_0}[f_i]\}_{i=1}^N$. 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 to find the α'-projection of p onto $\mathcal{M}$. Since it is given by $\arg\min_{p_0\in\mathcal{M}} D_{\alpha'}(p||p_0)$ (See Appendix for properties of α'-projection.), the α'-divergence between p and p_0 is expressed as,

$$\begin{aligned} D_{\alpha'}(p||p_0) &= \frac{4}{1-\alpha'^2}\left[1 - \sum_{\mathbf{x}} p^{(1-\alpha')/2}p'^{(1+\alpha')/2}\right] \\ &= \frac{4}{1-\alpha'^2}\left[1 - e^{-\frac{1-\alpha'}{2}\left(\psi + \frac{1+\alpha'}{2}G_{\alpha'}\right)}\right], \end{aligned} \tag{3}$$

where $\psi = \log Z$ and

$$\begin{aligned} \frac{1+\alpha'}{2}G_{\alpha'} &\equiv -\log Z + \frac{2}{\alpha'-1} \\ &\quad \times \log\sum_{\mathbf{x}} p^{(1-\alpha')/2}{p_0}^{(1+\alpha')/2}. \end{aligned} \tag{4}$$

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 $\arg\min_{p_0\in\mathcal{M}} D_{\alpha'}(p||p_0) = \arg\min_{p_0\in\mathcal{M}} G_{\alpha'}(p||p_0)$. We minimize $G_{\alpha'}$ instead of $D_{\alpha'}$ hereafter.

Then what α' should we choose to approximate $\boldsymbol{\eta}$ by $\boldsymbol{\eta}_0$? If $\mathcal{M}$ is a 1-autoparallel 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 α' taking account of the computational complexity.

Because p is a distribution of the α-family, it is represented as $p^{(1-\alpha)/2} = c_1 \sum_{\nu=0}^{\tilde{N}} \theta^\nu f_\nu + c_2$ with some constants c_1 and c_2. Suppose we choose α' such that the α'-divergence is a linear function of $p^{(1-\alpha)/2}$, we do not have to deal with any nonlinear functions of $\{f_\nu\}$ to calculate the value of $D_{\alpha'}(p||p_0)$. Because of this reason, we choose $\alpha = \alpha'$ here. In this case, $G_\alpha = -(4/(1-\alpha^2))\log(((1-\alpha)/2)\sum_{\nu=0}^{\tilde{N}} \theta^\nu \langle f_\nu\rangle_\alpha^0)$, with $\langle f_\nu\rangle_\alpha^0 \equiv \sum_{\mathbf{x}} f_\nu(\mathbf{x}) {p_0}^{(1+\alpha)/2}(\mathbf{x};\mathbf{h})$. If $f_\nu(\mathbf{x})$ $(\nu = 0, 1, \ldots, \tilde{N})$ are functions of l variables, it takes $O(\tilde{N}k^l)$ steps to calculate G_α. In addition, we can also calculate

$$\frac{\partial G_\alpha}{\partial h^i} = -\frac{4}{1-\alpha^2}\frac{\displaystyle\sum_\nu \theta^\nu \frac{\partial}{\partial h^i}\langle f_\nu\rangle_\alpha^0}{\displaystyle\sum_\nu \theta^\nu \langle f_\nu\rangle_\alpha^0} \tag{5}$$

for $i = 1, \ldots, N$ at the same orders. One can easily find $\mathbf{h}$ that gives a local minimum of G_α 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 $\boldsymbol{\eta}$ requires $O(Nk^N)$ operations in general.

3. Application to a Binary-State Model

In this section we calculate the α-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 $\sum_{\nu=1}^{\tilde{N}} \theta^\nu f_\nu(\mathbf{x}) = \sum_{i=1}^N \sum_{j>i} \theta^{ij} x_i x_j + \sum_{i=1}^N \theta^i x_i$. Here we assume that $\min_{\mathbf{x}}[((1-\alpha)/2)\sum_\nu \theta^\nu f_\nu] = c(>0)$. We approximate p by a p_0 that minimizes the α-divergence between them. Because $\mathbf{x}$ is a binary vector here, p_0 is generally represented in the following exponential form:

$$p_0 = \frac{1}{Z_0(\mathbf{h})}\exp\left(\sum_{i=1}^N h^i x_i\right). \tag{6}$$

Let us introduce the following expectations,

$$\begin{aligned} {\eta_0}_\nu^{(\alpha)} &\equiv \sum_{\mathbf{x}} f_\nu(\mathbf{x}) p_0\left(\mathbf{x}; \frac{1+\alpha}{2}\mathbf{h}\right) \\ &= \begin{cases} \tanh\left(\dfrac{1+\alpha}{2}h^i\right)\tan\left(\dfrac{1+\alpha}{2}h^j\right), & f_\nu = x_i x_j \\ \tanh\left(\dfrac{1+\alpha}{2}h^i\right), & f_\nu = x_i \end{cases} \end{aligned} \tag{7}$$

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

$$\frac{\partial G_\alpha}{\partial h^i} = \frac{2}{1-\alpha}\left[\eta_{0i} - {\eta_0}_i^{(\alpha)} - {g_0}_{ii}^{(\alpha)} \frac{\displaystyle\sum_{j=1}^N \theta^{ij} {\eta_0}_j^{(\alpha)} + \theta^i}{\displaystyle\sum_{\nu=1}^{\tilde{N}} \theta^\nu {\eta_0}_\nu^{(\alpha)}}\right] \tag{8}$$

with ${g_0}_{ii}^{(\alpha)} = [1 - ({\eta_0}_i^{(\alpha)})^2]$. We employ the gradient descent algorithm to search for a local minimum of G_α (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: $\eta_i = \tanh(\sum_j \theta^{ij}\eta_j + \theta^i)$ for $i = 1, \ldots, N$. 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 α-projection is a generalization of the naive mean-field method for α-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 1. Algorithm A.

1. Initialize $\mathbf{h}$ to small random values.
2. Calculate ${\eta_0}_\nu^{(\alpha)}$, ${g_0}_{i\nu}^{(\alpha)}$, and $\partial G_\alpha/\partial h^i$.
3. Update $\mathbf{h}$ according to
   $h^i_{\text{new}} = h^i_{\text{old}} - \delta\frac{\partial G_\alpha}{\partial h^i}$,
   where δ is a step size.
4. Return to step 2; stop after finite steps n^*.

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 $\eta_i = E_p[(p(x_i = 1, \mathbf{x}_{\setminus i}) - p(x_i = -1, \mathbf{x}_{\setminus i}))/(p(x_i = 1, \mathbf{x}_{\setminus i}) + p(x_i = -1, \mathbf{x}_{\setminus i}))]$, where $\mathbf{x}_{\setminus i} \equiv \{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_N\}$. Employing the naive mean-field approximation of the above equation, we obtain the self-consistent equations for $\boldsymbol{\eta}$:

$$\eta_i \approx \frac{p(x_i = 1, \boldsymbol{\eta}_{\setminus i}) - p(x_i = -1, \boldsymbol{\eta}_{\setminus i})}{p(x_i = 1, \boldsymbol{\eta}_{\setminus i}) + p(x_i = -1, \boldsymbol{\eta}_{\setminus i})}, \qquad (9)$$

for all i. We can compute the approximation of $\boldsymbol{\eta}$ 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.

Table 2. Algorithm B.

1. Initialize $\boldsymbol{\eta}_0$ to small random values.
2. Update $\boldsymbol{\eta}_0$ using Eq. (9).
3. Return to step 2; stop after finite steps n^{**}.

Figure 1 shows the average of normalized error: $\mathrm{Err} \equiv (\sum_{i=1}^N |\eta_i - \eta_{0i}| / \sum_{i=1}^N |\eta_i|)$ and its standard deviation obtained from Algorithms A (cross) and B (square) for $N = 10$ and $c = 1$. $\{\theta^{ij}\}$ and $\{\theta^i\}$ are generated from the Gauss distributions $\mathcal{N}(\mu, 0.5)$ and $\mathcal{N}(0.0, 1.0)$, respectively. We generate 100 samples here. Since it is generally difficult to find the step size δ of Algorithm A, we repeat this algorithm five times over different $\delta = \{1.0, 6.0, 11.0\}$ and choose the $\boldsymbol{\eta}_0$ that minimizes D_α 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 α

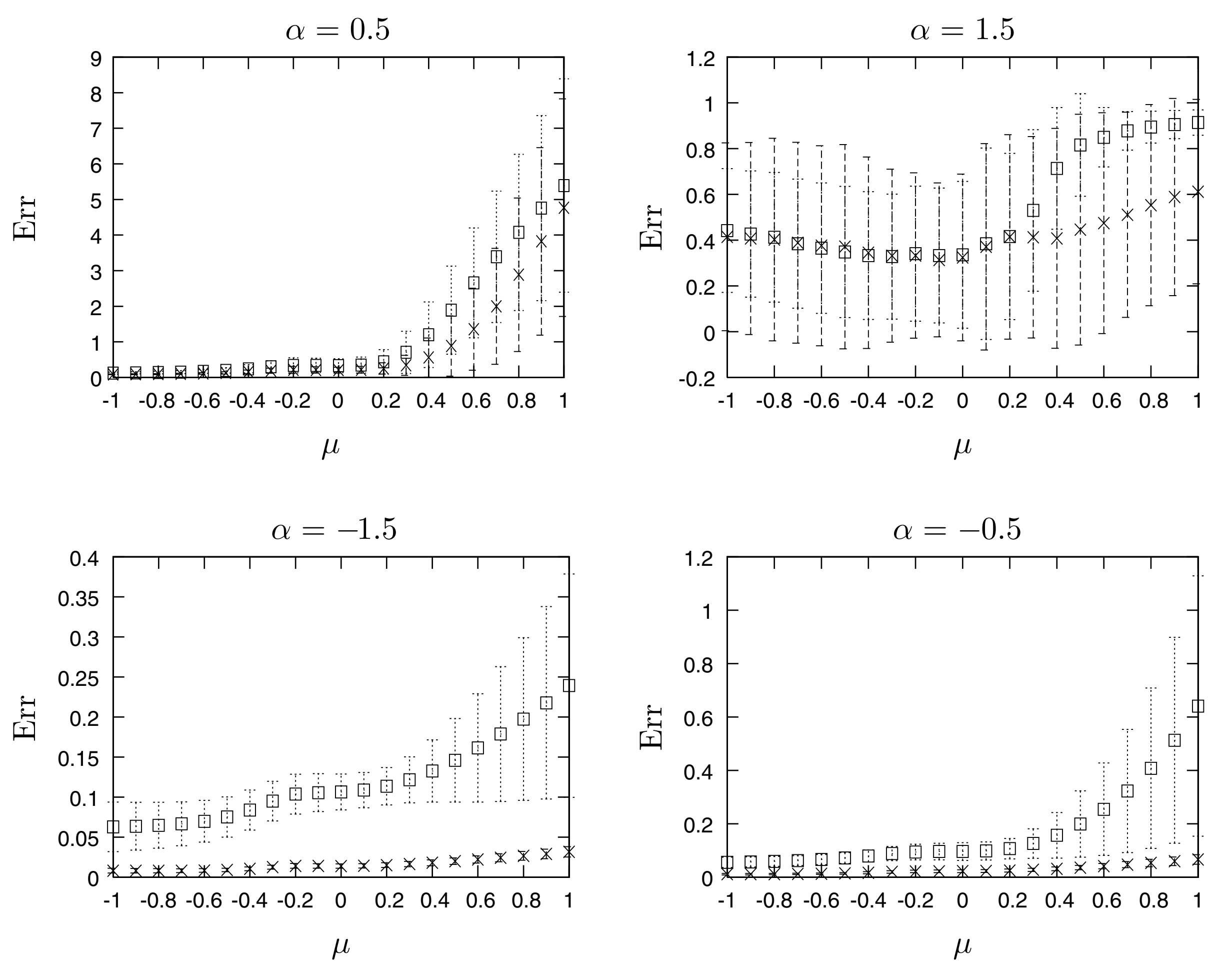

Fig. 1. The averaged normalized errors versus the mean of coupling coefficient μ.

that are shown here. If $\alpha = -1$, D_α have a unique minimum at which $\boldsymbol{\eta} = \boldsymbol{\eta}_0$ holds [Amari *et al.*, 2001]. Thus Algorithm A finds the exact expectations in this case, with a sufficiently small step size δ and many iterations. On the other hand, since the α-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 α' that most alleviate the computational cost to approximate $\boldsymbol{\eta}$. As we will see, there is a tradeoff between the computational complexity and the accuracy of the approximation.

If $p(\mathbf{x};\theta)$ 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 $\boldsymbol{\eta}$, we take here the simplest model to compare several projections.

Now we consider the α-family ($|\alpha| < 1$) of Eq. (1) with $\mathbf{x} = \{x_i | x_i \in \{+1, -1\}, i = 1, 2\}$ and $\mathcal{H}(\mathbf{x}) = \theta^{12} x_1 x_2 + \theta^1 x_1 + \theta^2 x_2$, where $\theta^{12} = (2/(1-\alpha))J$ and $\theta^1 = \theta^2 = (2/(1-\alpha))H$.

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

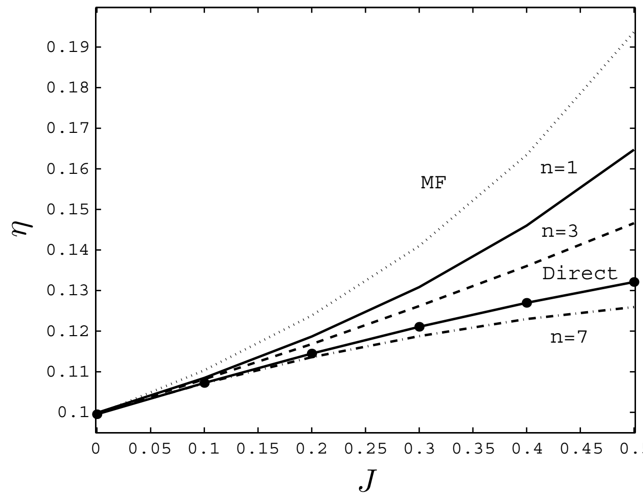

Fig. 2. Approximations of η versus J.

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 $\alpha'_n = 1-(1-\alpha)n$, we can also calculate the $G_{\alpha'_n}$ of Eq. (4) to be Eq. (15), where $\eta_0^{(\alpha')} = \tanh(((1+\alpha')/2)h)$, $(c_1, H_1, J_1) = (1, H, J)$, and

$$c_n = c_{n-1} + 2HH_{n-1} + JJ_{n-1}, \qquad (10)$$

$$H_n = H(c_{n-1} + J_{n-1}) + (1+J)H_{n-1}, \qquad (11)$$

$$J_n = J_{n-1} + c_{n-1}J + 2HH_{n-1}, \qquad (12)$$

for $n > 2$; we choose the η_0 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 $\alpha'_5 = -1$ in this case, the α'_n-projection gives a good approximation when $n \approx 5$.

$$\eta = \frac{[1+2H+J]^{2/(1-\alpha)} - [1-2H+J]^{2/(1-\alpha)}}{[1+2H+J]^{2/(1-\alpha)} + 2[1-J]^{2/(1-\alpha)} + [1-2H+J]^{2/(1-\alpha)}}, \qquad (13)$$

$$\eta_{\mathrm{MF}} = \frac{[1+H(\eta_{\mathrm{MF}}+1)+J\eta_{\mathrm{MF}}]^{2/(1-\alpha)} - [1+H(\eta_{\mathrm{MF}}-1)-J\eta_{\mathrm{MF}}]^{2/(1-\alpha)}}{[1+H(\eta_{\mathrm{MF}}+1)+J\eta_{\mathrm{MF}}]^{2/(1-\alpha)} + [1+H(\eta_{\mathrm{MF}}-1)-J\eta_{\mathrm{MF}}]^{2/(1-\alpha)}}, \qquad (14)$$

$$G_{\alpha'_n} = -\frac{4}{1-\alpha'^2}\left[\log(c_n + 2H_n\eta_0^{(\alpha')} + J_n(\eta_0^{(\alpha')})^2) - (1+\alpha')\log(2\cosh h) + 2\log\left(2\cosh\frac{1+\alpha'}{2}h\right)\right]. \qquad (15)$$

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

5. Conclusion

We have shown that it is possible to realize a generalization of the mean-field method by calculating the α-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 α-family and compared the method with the mean-field approximation of the Callen 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 α-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 α-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 Innovational 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

Abe, S. & Okamoto, Y. (eds.) [2001] *Nonextensive Statistical Mechanics and Its Application* (Springer-Verlag).

Amari, S. & Nagaoka, H. [2000] *Methods of Information Geometry* (The AMS & Oxford University Press).

Amari, S., Ikeda, S. & Shimokawa, H. [2001] "Information geometry of α-projection in mean field approximation," in *Advanced Mean Field Methods*, eds. Opper, M. & Saad, D. (MIT Press), pp. 241–257.

Arndt, C. [2001] *Information Measures* (Springer-Verlag).

Baldovin, F. & Rovledo, A. [2002] "Sensitivity to initial conditions at bifurcations in one-dimensional nonlinear maps: Rigorous nonextensive solutions," *Europhys. Lett.* **60**, 518–524.

Beck, C. [2001] "Dynamical foundations of nonextensive statistical mechanics," *Phys. Rev. Lett.* **87**.

Beck, C. [2002] "Generalized statistical mechanics and fully developed turbulence," *Physica A* **306**, 189–198.

Bhattacharyya, C. & Keerthi, S. S. [2000] "Information geometry and Plefka's mean-field theory," *J. Phys. A* **33**, 1307–1312.

Ghashghaie, S., Breymann, W., Peinke, J., Talkner, P. & Dodge, Y. [1996] "Turbulent cascades in foreign exchange markets," *Nature* **381**, 767–770.

Kabashima, Y. & Saad, D. [1998] "Belief propagation versus tap for decoding corrupted messages," *Europhys. Lett.* **44**, 668–674.

Latora, V., Rapisarda, A. & Ruffo, S. [1998] "Lyapnov instability and finite size effects in a system with long-range forces," *Phys. Rev. Lett.* **80**, 692–695.

Latora, V. & Tsallis, C. [2001] "Fingerprints of nonextensive thermodynamics in a long-range Hamiltonian system," *Physica A* **305**, 129–136.

Latora, V., Rapisarda, A. & Tsallis, C. [2001] "Non-gaussian equilibrium in a long-range Hamiltonian system," *Phys. Rev. E* **64**.

Lenzi, E. K., Malacarne, L. C. & Mendes, R. S. [1998] "Perturbation and variational methods in nonextensive Tsallis statistics," *Phys. Rev. Lett.* **80**, 218–221.

Mendes, R. S., Kwok, S. F., Lenzi, E. K. & Maki, J. N. [1999] "Perturbation expansion, Bogolyubov inequality and integral representations in nonextensive Tsallis statistics," *Europ. Phys. J. B* **10**, 353–359.

Opper, M. & Winther, O. [2000] "Gaussian processes for classification: Mean field algorithms," *Neural Comput.* **12**, 2655–2684.

Peterson, C. & Anderson, J. R. [1987] "A mean field theory learning algorithm for neural networks," *Compl. Syst.* **1**, 995–1019.

Plastino, A. & Tsallis, C. [1993] "Variational method in generalized statistical mechanics," *J. Phys. A* **26**, L893–L896.

Sarmento, E. F. [1995] "Generalization of the single-site callen identity within Tsallis statistics," *Physica A* **218**, 482–486.

Tanaka, T. [2000] "Information geometry of mean-field approximation," *Neural Comput.* **12**, 1951–1968.

Taruya, A. & Sakagami, M. [2002] "Gravothermal catastrophe and Tsallis' generalized entropy of self-gravitating systems," *Physica A* **307**, 185–206.

Toyoizumi, T. & Aihara, K. [2003] "Mean-field and variational methods for α-families," *Trans. IEICE D-II* (*in Japanese*) **J86-D-II**, 959–965.

Tsallis, C. [1988] "Possible generalization of Boltzmann-Gibbs statistics," *J. Statist. Phys.* **52**, 479–487.

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 $\mathcal{S} = \{p(\mathbf{x};\boldsymbol{\theta})\}$ be an α-family. The α-divergence between two probability distributions $p = p(\mathbf{x};\boldsymbol{\theta})$ and $p' = p(\mathbf{x};\boldsymbol{\theta}')$ is defined as

$$D_\alpha(p||p') \equiv \frac{4}{1-\alpha^2}\left[1 - \sum_{\mathbf{x}} p^{(1-\alpha)/2} p'^{(1+\alpha)/2}\right] \tag{A.1}$$

for $\alpha \neq \pm 1$, and $D_{\pm 1}(p||p') \equiv \lim_{\alpha\to\pm 1} D_\alpha(p||p')$ for $\alpha = \pm 1$. Note that D_{-1} is equivalent to the well-known Kullback–Leibler divergence. The non-negative property of the α-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 α-divergence is a distance-like measure representing the difference between two probability distributions.

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

Theorem 1. *Let $\mathcal{S}$ be an α-family, p, q and r be three points in $\mathcal{S}$, γ_1 be an α-geodesic connecting p and q and γ_2 be an $(-\alpha)$-geodesic connecting q and r in $\mathcal{S}$. If the curves γ_1 and γ_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 $\mathcal{S}$ be an α-family, $\mathcal{M}$ a submanifold of $\mathcal{S}$, and p a point in $\mathcal{S}$. A necessary and sufficient condition for a point $q \in \mathcal{M}$ to be a stationary point of the function $r \mapsto D_\alpha(p||r)$ restricted on $\mathcal{M}$ is that the α-geodesic connecting p and q to be orthogonal to $\mathcal{M}$ at q. We call such q as an α-projection of p onto $\mathcal{M}$.*