Duality in a maximum generalized entropy model

Shinto Eguchi*, Osamu Komori* and Atsumi Ohara†

*Institute of Statistical Mathematics, Japan
†University of Fukui, Japan

Abstract. This paper discusses a possible generalization for the maximum entropy principle. A class of generalized entropy is introduced by that of generator functions, in which the maximum generalized distribution model is explicitly derived including q-Gaussian distributions, Wigner semicircle distributions and Pareto distributions. We define a totally geodesic subspace in the total space of all probability density functions in a framework of information geometry. The model of maximum generalized entropy distributions is shown to be totally geodesic. The duality of the model and the estimation in the maximum generalized principle is elucidated to give intrinsic understandings from the point of information geometry.

Keywords: β-divergence, Dual connections, Generalized entropy, Generalized divergence, Information geometry

INTRODUCTION

The maximum entropy method consists of a statistical modeling and estimation based on the Boltzmann-Gibbs-Shannon entropy

$$H_{\rm BGS}(f) = -\int f(x)\log f(x) d\Lambda(x),$$

for a probability density function $f(x)$ with respect to a carrier measure Λ. Let $(X_1,\cdots,X_n)$ be a random sample from $f(x)$ and $t(x)$ be a feature vector. Then we consider a mean equal space for $t(X)$ as

$$\Gamma(\hat{t}) = \{f \in \mathscr{F} : \mathbb{E}_f\{t(X)\} = \hat{t}\},$$

where $\mathscr{F}$ is the space of all probability density functions and $\hat{t}$ is the sample mean vector, $\hat{t} = \sum_{i=1}^n t(X_i)/n$. The statistical model of maximum entropy distributions under the constraint $\Gamma(\hat{t})$ is characterized by an exponential model

$$f_{\rm exp}(x,\theta) = \exp\{\theta^\top t(x) - \kappa(\theta)\},$$

where $\kappa(\theta) = \log\int \exp\{\theta^\top t(x)\} d\Lambda(x)$. Thus the estimator $\hat{\theta}$ for θ is given by the mean matching

$$\mathbb{E}_{f_{\rm exp}(\cdot,\hat{\theta})}\{t(X)\} = \hat{t},$$

which is equal to the likelihood equation, so that $\hat{\theta}$ is nothing but the maximum likelihood estimator. The class of model includes Gaussian distributions, Poisson distributions and Gibbs distributions. The maximum entropy method has been widely employed in fields such as natural language processing [4], ecological analysis [23] and so forth.

On the other hand, there is another type of entropy measure such as the Hill diversity index, the Gini-Simpson index, the Tsallis entropy and so on, cf. [26, 15, 27] from different fields. We introduced the class of generalized entropy measures to include all the entropy measures mentioned above. We discuss the maximum generalized entropy principle in this extension of the Boltzmann-Gibbs-Shannon entropy. The model of generalized maximum entropy distributions includes q-Gaussian distributions, Wigner distributions and Pareto distributions. The estimation is given by minimum divergence method in which the divergence is led from the generalized entropy.

GENERALIZED ENTROPY

We introduce a class of generalized entropy that is constructed by a generator function U, see [8]. The class of generator functions is defined by

$$\mathscr{U} = \{U : \mathbb{R} \to \mathbb{R}_+ : U'(s) \geq 0, U''(s) \geq 0, U'''(s) \geq 0\}. \tag{1}$$

Then we consider the conjugate convex function defined on $\mathbb{R}_+$ of U in $\mathscr{U}$ as

$$U^*(t) = \max_{s \in \mathbb{R}} \{st - U(s)\}, \tag{2}$$

and hence $U^*(t) = tu^{-1}(t) - U(u^{-1}(t))$, where $u(s) = U'(s)$. Note that there exists the inverse function of $u(t)$ in (2) since U is assumed to be in $\mathscr{U}$. We define a generalized entropy

$$H_U(f) = -\int U^*(f) d\Lambda, \tag{3}$$

which is called U-diagonal entropy. Similarly, the U-cross entropy is given by

$$C_U(f,g) = \int \{U(u^{-1}(g)) - fu^{-1}(g)\} d\Lambda,$$

and hence $H_U(f) = C_U(f,f)$. The information divergence

$$D_U(f,g) = C_U(f,g) - H_U(f), \tag{4}$$

called U-divergence. We note from the assumption of $U \in \mathscr{U}$ that the $D_U(f,g) \geq 0$ with equality if and only if $g = f$ in Λ-everywhere.

The most typical example of U is $U_0(s) = \exp(s)$, which leads to $U_0^*(t) = t\log t - t$. Thus U_0-divergence and U_0-entropy equal the Kullback-Leibler divergence and the Boltzmann-Gibbs-Shanon entropy, respectively. As a further example consider the function

$$U_\beta(s) = \frac{1}{\beta+1}(1+\beta s)^{\frac{1+\beta}{\beta}} \tag{5}$$

where $\beta < 1$ is a scalar. Then the generator function U_β associates with the β-diagonal power entropy

$$H_\beta(f) = -\frac{1}{\beta(\beta+1)} \int f(x)^{\beta+1} d\Lambda(x) + \frac{1}{\beta},$$

the β-power cross entropy

$$C_\beta(f,g) = \int \Big\{ \frac{1}{\beta+1} g(x)^{\beta+1} - \frac{1}{\beta} f(x)\{g(x)^\beta - 1\} \Big\} d\Lambda(x)$$

and the β-power divergence

$$D_\beta(f,g) = \frac{1}{\beta(\beta+1)} \int \Big\{ f(x)^{\beta+1} - (\beta+1) f(x) g(x)^\beta + \beta g(x)^{\beta+1} \Big\} d\Lambda(x).$$

We observe that the β-power entropy reduces to the Boltzmann-Gibbs-Shannon entropy in the limit of β to 0 and similarly the β-power divergence reduces to the Kullback-Leibler divergence. If we take a limit of β to -1, then $D_\beta(f,g)$ becomes the Itakura-Saito divergence

$$D_{\mathrm{IS}}(f,g) = \int \Big\{ -\log \frac{f(x)}{g(x)} + \frac{f(x)}{g(x)} - 1 \Big\} d\Lambda(x),$$

which is widely applied in signal processing and speech recognition, cf. [25, 5].

The β-power entropy H_β is essentially equal to the Tsallis q-entropy with a relation $q = \beta + 1$, cf. [27, 19, 28]. Tsallis entropy has essential understandings for phenomena of spin glass relaxation, dissipative optical lattices and so on beyond the classical statistical physics associated with the Boltzmann-Shannon entropy $H_0(p)$. See also [26, 15] for the power entropy in the field of ecology. The statistical property for the minimum β divergence method in the presence of outliers departing from a supposed model is discussed to show a robustness performance by appropriate selection for β, cf. [17, 12, 13] and a property of spontaneous learning to apply to clustering analysis is focused beyond robustness perspective as in [21].

MAXIMUM GENERALIZED ENTROPY MODEL

We discuss the principle of maximum generalized entropy. In general the U-entropy is an unbounded functional on $\mathscr{F}$ unless $\mathscr{F}$ is of finite discrete case. For this we introduce a moment constraint as follows. Let $t(X)$ be a k-dimensional statistic vector. Henceforth we consider the mean equal space $\Gamma(\tau)$ as in Introduction assuming that $\mathbb{E}_f\{\|t(X)\|^2\} < \infty$ for all f of $\mathscr{F}$.

Theorem 1. *Let* $f_\tau{}^* = \mathrm{argmax}\{H_U(f) : f \in \Gamma(\tau)\}$, *where* $H_U(f)$ *is* U*-diagonal entropy defined in* (3). *Then the maximum* U*-entropy distribution is given by*

$$f_\tau{}^*(x) = u(\theta^\top t(x) - \kappa_U(\theta)), \tag{6}$$

where $\kappa_U(\theta)$ is the normalizing factor and θ is a parameter vector determined by the moment constraint

$$\int t(x)u(\theta^\top t(x)-\kappa_U(\theta))d\Lambda(x)=\tau.$$

Proof. For any $f_\tau(x)$ in $\Gamma(\tau)$ we observe that

$$\mathbb{E}_{f_\tau}\{u^{-1}(f_\tau{}^*(X))\}=\mathbb{E}_{f_\tau{}^*}\{u^{-1}(f_\tau{}^*(X))\}$$

Therefore we can confirm that $H_U(f_\tau{}^*)\geq H_U(f_\tau)$ for any $f_\tau\in\Gamma(\tau)$ since

$$H_U(f_\tau{}^*)-H_U(f_\tau)=D_U(f_\tau,f_\tau{}^*),$$

which is nonnegative by the definition of U-divergence. The proof is complete.

Here we give a definition of the model of maximum U-entropy distributions as follows.

Definition 1. *We define a k-dimensional model*

$$M_U=\{f_U(x,\theta):=u(\theta^\top t(x)-\kappa_U(\theta)):\theta\in\Theta\},\tag{7}$$

which is called U-model, where $\Theta=\{\theta\in\mathbb{R}^k:\kappa_U(\theta)<\infty\}$.

The Naudts' deformed exponential family discussed from a statistical physical viewpoint as in [19] is closely related with U-model. We discuss a typical example by the power entropy $H_\beta(f)$, see [18, 19] from a viewpoint of statistical physics. Consider a mean equal space of univariate distributions on $\mathbb{R}_+$

$$\Gamma(\mu,\sigma^2)=\{f:\mathbb{E}_f(X)=\mu,\mathbb{V}_f(X)=\sigma^2\}.$$

The maximum entropy distribution with H_β is given by

$$f_\beta(x,\mu,\sigma^2)=\frac{1}{\sigma}\Big(1-\frac{\beta}{1+\beta}\frac{(x-\mu)}{\sigma}\Big)_+^{\frac{1}{\beta}},$$

which is nothing but Pareto distribution. We next consider a case of multivariate distributions, where the moment constraints are supposed that for a fixed p-dimensional vector μ and matrix V of size $p\times p$

$$\Gamma(\mu,V)=\{f\in\mathscr{F}:\mathbb{E}_f(X)=\mu,\mathbb{V}_f(X)=V\}.$$

Let $f_\beta(\cdot,\mu,V)=\mathrm{argmax}_{f\in\Gamma(\mu,V)}H_\beta(f)$. If we consider a limit case of β to 0, then $H_\beta(f)$ reduces to $H_{\mathrm{BGS}}(f)$ and the maximum entropy distribution is a p-dimensional Gaussian distribution with the density function

$$\varphi(x,\mu,V)=\{\det(2\pi V)\}^{-p/2}\exp\Big\{-\frac{1}{2}(x-\mu)^\top V^{-1}(x-\mu)\Big\}.$$

In general we deduce that if $\beta > -2/(p+2)$, then the maximum β-power entropy distribution uniquely exists such that the density function is given by

$$f_\beta(x,\mu,V) = \frac{c_\beta}{\det(2\pi V)^{\frac{1}{2}}}\Big\{1 - \frac{\beta}{2+p\beta+2\beta}(x-\mu)^\top V^{-1}(x-\mu)\Big\}_+^{\frac{1}{\beta}},$$

where c_β is the normalizing factor, see [9, 10] for the detailed expression and [22] for the group invariance perspective. If $\beta > 0$, then the maximum β-power entropy distribution has a compact support, in which the typical case is $\beta = 2$ called the Wigner semicircle distribution. On the other hand, if $-2/(p+2) < \beta < 0$, the maximum β-power entropy distribution has a full support of $\mathbb{R}^p$, and equals a p-variate t-distribution with a degree of freedom depending on β.

MINIMUM DIVERGENCE METHOD

We consider a general situation where the underlying density function $f(x)$ is sufficiently approximated by a statistical model $M = \{f(x,\theta) : \theta \in \Theta\}$. The U-loss function for a given data set $\{X_i : i = 1,\cdots,n\}$ is introduced by

$$L_U(\theta) = -\frac{1}{n}\sum_{i=1}^n u^{-1}\big(f(X_i,\theta)\big) + b_U(\theta),$$

where $b_U(\theta) = \int U\big(u^{-1}\big(f(x,\theta)\big)\big)d\Lambda(x)$. We call $\hat{\theta}_U = \mathrm{argmin}_{\theta\in\Theta} L_U(\theta)$ U-estimator for the parameter θ. By definition $\mathbb{E}_f\{L_U(\theta)\} = C_U(f, f(\cdot,\theta))$ for all θ in Θ, which implies that $L_U(\theta)$ almost surely converges to $C_U(f, f(\cdot,\theta))$ as n goes to ∞. Let us define a statistical functional as

$$\theta_U(f) = \operatorname*{argmin}_{\theta\in\Theta} C_U(f, f(\cdot,\theta)).$$

Then $\theta_U(f)$ is model-consistent, or $\theta_U(f(\cdot,\theta)) = \theta$ for any $\theta \in \Theta$ because

$$C_U(f(\cdot,\theta), f(\cdot,\theta')) \le H_U(f(\cdot,\theta))$$

with equality if and only if $\theta' = \theta$. Hence U-estimator $\hat{\theta}_U$ is asymptotically consistent. There is a natural question which situation happens if we consider the U-estimation under the U-model?

Let M_U be a U-model defined in (7). Then the U-loss function under the U-model for a given data set $\{X_1,\cdots,X_n\}$ is defined by

$$L_U(\theta) = -\theta^\top \hat{t} + \kappa_U(\theta) + b_U(\theta), \tag{8}$$

where $\hat{t} = \sum_{i=1}^n t(X_i)/n$ and $b_U(\theta) = \int U(u^{-1}(\theta^\top t(x) - \kappa_U(\theta))d\Lambda(x)$. The estimating equation is given by

$$\frac{\partial}{\partial\theta}L_U(\theta) = -\hat{t} + \mathbb{E}_{f(\cdot,\theta)}\{t(X)\}.$$

Hence, if we consider the U-estimator for a parameter η by the transformation of θ defined by $\varphi(\theta) = \mathbb{E}_{f(\cdot,\theta)}\{t(X)\}$, then the U-estimator $\hat{\eta}_U$ is nothing but the sample mean $\hat{t}$. Here we observe that the transformation $\varphi(\theta)$ is one-to-one. Consequently the estimator $\hat{\theta}_U$ for θ is given by $\varphi^{-1}(\hat{t})$. We summarize theses results as follows.

Theorem 2. *Let M_U be a U-model with a canonical statistic $t(X)$ as defined in (7). Then the U-estimator for the expectation parameter η of $t(X)$ is always the sample mean $\hat{t}$.*

We remark that the empirical Pythagorean theorem holds as in

$$L_U(\theta) = L_U(\hat{\theta}_U) + D_U(\hat{\theta}_U, \theta),$$

since we observe that

$$L_U(\theta) - L_U(\hat{\theta}_U) = (\hat{\theta}_U - \theta)^\top \hat{t} + \kappa_U(\theta) + b_U(\theta) - \kappa_U(\hat{\theta}_U) + b_U(\hat{\theta}_U),$$

which gives another proof for which $\hat{\theta}_U$ is $\varphi^{-1}(\hat{t})$. The statistic $\hat{t}$ is a sufficient statistic in the sense that the U-loss function $L_U(\theta)$ is a function of $\hat{t}$ as in (8). Accordingly the U-estimator under U-model is a function only of $\hat{t}$ from the observations $X_1, \cdots, X_n$. This is an extension that the MLE is a function of $\hat{t}$ under the exponential model with the canonical statistic $t(X)$.

Let us look at the case of the β-power divergence. Under the β-power model given by

$$M_\beta = \{ f_\beta(x,\theta) := \{\kappa_\beta(\theta) + \beta\theta^\top t(x)\}^{\frac{1}{\beta}} : \theta \in \Theta\},$$

the β-loss function is written by

$$L_\beta(\theta) = -\beta\theta^\top \hat{t} + \kappa_\beta(\theta) + b_\beta(\theta),$$

where

$$b_\beta(\theta) = \frac{1}{\beta+1}\int \{\kappa_\beta(\theta) + \beta\theta^\top t(x)\}^{\frac{1+\beta}{\beta}} d\Lambda(x).$$

The β-power estimator for the expectation parameter of $t(X)$ is exactly given by $\hat{t}$.

DUALITY

We discuss duality in a maximum generalized entropy model. For this we introduce a path geometry in the space $\mathscr{F}$ of all density functions, see the framework of information geometry, cf. [1, 2]. In particular the nonparametric formulation is discussed in [24, 29, 3, 11]. Let f and g be in $\mathscr{F}$ and ϕ be a strictly-increasing and convex function defined in $\mathbb{R}_+$. Then we call

$$C^{\phi} = \{ f_t^{(\phi)} := \phi\big((1-t)\phi^{-1}(f) + t\phi^{-1}(g) - \kappa_t(f,g)\big) : 0 \le t \le 1\} \tag{9}$$

ϕ-geodesic connecting with f with g, where $\kappa_t(f,g)$ is a normalizing factor to satisfy $\int f_t^{(\phi)} d\Lambda = 1$. Note from the convexity assumption for ϕ that the Nagumo-Kolmogorov average satisfies

$$\phi((1-t)\phi^{-1}(f)+t\phi^{-1}(g)) \leq (1-t)f+tg$$

for all t, $0 \leq t \leq 1$. This guarantees the existence of $\kappa_t(f,g)$. This definition is an extension of mixture geodesic curve $C^{(\mathrm{m})} = (1-t)f+tg$, which is a special choice of $\phi = \mathrm{id}$.

Let M be a submanifold of $\mathscr{F}$. We say M is totally ϕ-geodesic if the ϕ-geodesic curve defined in (9) is embedded in M for any f and g in M. By definition the mean equal space $\Gamma(\tau)$ is totally mixture geodesic, that is, if f and g are in $\Gamma(\tau)$, then $(1-t)f+tg$ is also in $\Gamma(\tau)$ for any $t \in (0,1)$. We have a geometric understanding for the U-model similar to the exponential model.

Theorem 3. *Let M_U be a statistical model defined in* (7)*, where U is in $\mathscr{U}$ defined in* (1)*. Then M_U is totally ϕ-geodesic, where $\phi = u$.*

Proof. For arbitrarily fixed θ_1 and θ_2 in Θ, we observe that

$$f_t^{(\phi)} = \phi\big((1-t)u^{-1}(f_U(\cdot,\theta_1))+tu^{-1}(f_U(\cdot,\theta_2))-\kappa_t(\theta_1,\theta_2)\big)$$

with a normalizing factor $\kappa_t(\theta_1,\theta_2)$. Hence we conclude that, if $\phi = u$, then $f_t^{(\phi)} = f_U(\cdot,\theta_t)$, where $\theta_t = (1-t)\theta_1+t\theta_2$. We see from the convexity of Θ that $\theta_t \in \Theta$ for all $t, 0 \leq t \leq 1$, where Θ is defined in Definition 1. This completes the proof.

Hence the total space $\mathscr{F}$ is decomposed into $\Gamma(\tau)$ and M_U, where the intersection of $\Gamma(\tau)$ and M_U is a singlton of $f_U(\cdot,\theta)$ satisfying

$$\mathbb{E}_{f_U(\cdot,\theta)}\{t(X)\} = \tau.$$

The decomposition of $\mathscr{F}$ forms a foliation

$$\mathscr{F} = \bigcup_{f \in M_U} \Gamma(\tau_f),$$

where $\tau_f = \mathbb{E}_f\{t(X)\}$. In the foliation $\Gamma(\tau_f)$ is totally mixture geodesic; M_U is totally u-geodesic. From a differential geometry associated the U-divergence the dual connections are formulated in [6, 7, 11]. In fact the two connections leads to mixture geodesic and u-geodesic. We can say that the mixture geodesic and u-geodesic are dual in the sense that the average of two connections is the Levi-Civita connection with respect to the Riemannian metric associated with the U-divergence.

REFERENCES

1. Amari, S. *Differential-geometrical methods in statistics,* Lecture Notes in Statist., 28, Springer, New York, 1985.

2. Amari, S. and Nagaoka, H. *Methods of Information Geometry*. Oxford University Press, Oxford, UK, 2000.
3. Amari, S-I. Information Geometry of Positive Measures and Positive-Definite Matrices: Decomposable Dually Flat Structure. *Entropy* 2014, *16*, 2131-2145.
4. Berger, A. L., Pietra, V. J. D., and Pietra, S. A. D. A maximum entropy approach to natural language processing. *Computational linguistics* 1996, *22*, 39-71.
5. Cichocki, A. and Amari, S. I. Families of alpha-beta-and gamma-divergences: Flexible and robust measures of similarities. *Entropy* 2010, *12*, 1532-1568.
6. Eguchi, S. Second order efficiency of minimum contrast estimators in a curved exponential family. *Annals of Statistics* 1983, *11*, 793-803.
7. Eguchi, S. Geometry of minimum contrast. *Hiroshima Math. J* 1992, *22*, 631-647.
8. Eguchi, S. Information divergence geometry and the application to statistical machine learning. In *Information Theory and Statistical Learning*, 309-332. Eds. F. Emmert-Streib and M. Dehmer, Springer US, 2008.
9. Eguchi, S. and Kato, S. Entropy and divergence associated with power function and the statistical application. *Entropy* 2010, *12*, 262-274.
10. Eguchi, S., Komori, O. and Kato, S.; Projective Power Entropy and Maximum Tsallis Entropy Distributions. *Entropy* 2011, *13* 1746-1764.
11. Eguchi, S., Komori, O. and Ohara, A.; Duality of maximum entropy and minimum divergence. *Entropy* (2014) 16, 7, 3552-3572. .
12. Fujisawa, H. and Eguchi, S. Robust estimation in the normal mixture model. *J. Statist. Plan. Infer.* 2006, *136*, 3989-4011. .
13. Fujisawa, H. and Eguchi, S. Robust parameter estimation with a small bias against heavy contamination. *J. Multivariate Anal.* 2008, *99*, 2053-2081.
14. Grunwald, P. D., and Dawid, A. P. Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *Annals of Statistics* 2004, *32s*, 1367-1433.
15. Hill, M. O. Diversity and evenness: a unifying notation and its consequences. *Ecology 54*, 1973, 427–432.
16. Jaynes, E. T. Information Theory and Statistical Mechanics in *Statistical Physics*, K. Ford (ed.), Benjamin, New York, 1963.
17. Minami, M. and S. Eguchi. Robust blind source separation by beta divergence. *Neural computation* 2002, *14*, 1859-1886.
18. Naudts, J. The *q*-exponential family in statistical Physics. *Central European Journal of Physics* 2009, *7*, 405-413.
19. Naudts, J. *Generalized thermostatistics*, Springer, 2011.
20. Ohara, A.; Eguchi, S. Geometry on positive definite matrices deformed by V-potentials and Its submanifold structure, *Geometric Theory of Information* F. Nielsen *eds.*, Chapter 2 pp.31-55, Springer 2014.
21. A. Notsu, O. Komori and S. Eguchi. Spontaneous clustering via minimum gamma-divergence. *Neural Computation* 2014, *26* , 421-448.
22. Ohara, A. and Eguchi, S. Group invariance of information geometry on q-Gaussian distributions induced by beta-divergence. *Entropy* 2013, *15*, 4732-4747.
23. Phillips, S. J., and Dudik, M. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography* 2008, *31*, 161-175.
24. Pistone, G. and Sempi, C. An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. *Annals of Statistics* 1995,, 1543-1561.
25. Scharf, L. L. *Statistical signal processing*. Vol. 98. Reading, MA: Addison-Wesley, 1991.
26. Simpson, E. H. Measurement of diversity. *Nature*, *163*, 1949, 688.
27. Tsallis, C. Possible generalization of Boltzmann-Gibbs statistics. *J. Statistical Physics* 1988, *52*, 479-487.
28. Tsallis, C. *Introduction to Nonextensive Statistical Mechanics*; Springer, New York NY, USA, 2009.
29. Zhang, J. Nonparametric information geometry: From divergence function to referential-representational biduality on Statistical Manifolds. *Entropy* 2013, *15*, 5384-5418.