

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/256994855>

Discrete approximations of continuous distributions by maximum entropy

Article in *Economics Letters* · March 2013

DOI: 10.1016/j.econlet.2012.12.020

CITATIONS

14

READS

602

2 authors:



[Ken'ichiro Tanaka](#)

The University of Tokyo

42 PUBLICATIONS 212 CITATIONS

[SEE PROFILE](#)



[Alexis Akira Toda](#)

University of California, San Diego

65 PUBLICATIONS 308 CITATIONS

[SEE PROFILE](#)



Discrete approximations of continuous distributions by maximum entropy

Ken'ichiro Tanaka^a, Alexis Akira Toda^{b,*}

^a School of Systems Information Science, Future University Hakodate, Japan

^b Department of Economics, Yale University, United States

ARTICLE INFO

Article history:

Received 18 July 2012

Received in revised form

5 December 2012

Accepted 14 December 2012

Available online 22 December 2012

JEL classification:

C63

C65

Keywords:

Discrete approximation

Fenchel duality

Kullback–Leibler information

Maximum entropy principle

ABSTRACT

In numerically implementing the optimization of an expected value in many economic models, it is often necessary to approximate a given continuous probability distribution by a discrete distribution. We propose an approximation method based on the principle of maximum entropy and minimum Kullback–Leibler information, which is computationally very simple. Our method is not intended to replace existing methods but to complement them by “fine-tuning” probabilities so as to match prescribed (not necessarily polynomial) moments exactly.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

How can we find a discrete distribution that approximates a given distribution?

To motivate this question, consider the following problem. There is an investor who lives for two periods with utility function

$$\frac{1}{1-\gamma} \left(c_0^{1-\gamma} + \beta E[c_1^{1-\gamma}] \right), \quad (1)$$

where $\beta > 0$ is the discount factor, $\gamma > 0$ is the relative risk aversion coefficient, and c_0, c_1 are consumption for today and tomorrow. The investor is endowed with initial wealth $w > 0$ today but nothing tomorrow. After deciding how much to consume today, he can invest his remaining wealth $w - c_0$ in J assets indexed by $j = 1, \dots, J$. Asset j has gross return $R_j \geq 0$, which is a random variable. Let θ_j be the fraction of the remaining wealth invested in asset j , $\theta = (\theta_1, \dots, \theta_J) \in \mathbb{R}^J$ (where $\sum_j \theta_j = 1$) be the portfolio, and $R(\theta) = \sum_j R_j \theta_j$ be the return on portfolio θ . The budget constraint is then

$$c_1 = R(\theta)(w - c_0). \quad (2)$$

The investor's objective is to maximize the expected utility (1) subject to the budget constraint (2).

Characterizing the solution to this problem is not difficult. Substituting (2) into (1), it suffices to solve

$$\max_{c, \theta} \frac{1}{1-\gamma} \left(c^{1-\gamma} + \beta E[R(\theta)^{1-\gamma}] (w - c)^{1-\gamma} \right),$$

which can be broken into

$$\frac{k}{1-\gamma} := \max_{\theta} \frac{1}{1-\gamma} E[R(\theta)^{1-\gamma}], \quad (3a)$$

$$U := \max_c \frac{1}{1-\gamma} (c^{1-\gamma} + \beta k (w - c)^{1-\gamma}). \quad (3b)$$

Since the optimal consumption/saving problem (3b) can be solved by calculus, the original problem reduces to solving the optimal portfolio problem (3a).

Solving this problem numerically is not trivial, however, since for most probability distributions the expectation $E[R(\theta)^{1-\gamma}]$ admits no closed-form expression. However, if the distribution of asset returns $\mathbf{R} = (R_1, \dots, R_J)$ is discrete, then $E[R(\theta)^{1-\gamma}]$ is simply a sum and we can apply any optimization routine. Therefore, in practice it is important to approximate a given probability distribution by a discrete distribution.

A typical procedure for approximating a probability distribution by a discrete distribution is to partition the range of possible values (Tauchen, 1986) or the range of cumulative probabilities (Adda and Cooper, 2003) into intervals, and then assign the true probability to a representative point (usually the mid-point or the median) of each interval. The advantage of these methods is their simplicity

* Corresponding author. Tel.: +1 203 432 3576.

E-mail addresses: ketanaka@fun.ac.jp (K. Tanaka), alexisakira.toda@yale.edu (A.A. Toda).

¹ Of course, the investor is long in asset j if $\theta_j > 0$ and short if $\theta_j < 0$.

and that they work in any dimension. As shown by Miller and Rice (1983), however, this method underestimates the moments of the true distribution. Although the approximated moments approach the true moments as the number of points increases, we cannot always use a large number of points due to the computational cost with the problem we want to solve by discretization (in our case, (3a)). With a small number of points the discrepancy in moments can be substantial. As a remedy, Miller and Rice (1983) propose an approximation method based on Gaussian quadrature that match prescribed moments exactly, but their method works only in one dimension and with polynomial moments. DeVuyst and Preckel (2007) generalize the Gaussian quadrature method (which they call Gaussian cubature) to the multi-dimensional case, which reduces to solving a linear programming problem. However, their method is computationally intensive because the number of unknown variables is equal to that of points, which is typically large, and they do not prove that a solution exists.

In this paper we propose an approximation method based on Jaynes (1957)'s maximum entropy principle (MaxEnt) that matches prescribed moments exactly. Starting from any integration (quadrature) formula, we “fine-tune” the prior probabilities by minimizing the Kullback–Leibler information between the prior and posterior probabilities subject to the prescribed moment constraints. Since the dual problem of finding the probabilities is an unconstrained convex optimization problem with the number of unknown variables equal to the number of moments prescribed, it is computationally very simple. Furthermore, our method works on any discrete set (not necessarily a lattice) of any dimension with any prescribed moments (not necessarily polynomials).

2. Main results

2.1. Formulation of the problem

A researcher wants to approximate a probability density function f on $\mathbb{R}^K$ by probabilities $\{p(x)|x \in D\}$ on a finite discrete set $D \subset \mathbb{R}^K$. The density f may be known or unknown. For instance, f may represent a parametric distribution exogenously specified in the economic model, a nonparametric distribution estimated from data, or a prior distribution in the researcher's mind. In either case, we assume that some moments $\bar{T} = \int T(x)f(x)dx$ are given, where $T: \mathbb{R}^K \rightarrow \mathbb{R}^L$ is a measurable function. For instance, if the first and second moments are given,

$$T(x) = (x_1, \dots, x_K, x_1^2, \dots, x_K x_L, \dots, x_K^2).$$

Since we are prescribed K expectations, K variances, and $\frac{K(K-1)}{2}$ covariances, in this case we have

$$L = K + K + \frac{K(K-1)}{2} = \frac{K(K+3)}{2}.$$

To match these moments with a discrete distribution, it suffices to assign probabilities $\{p(x)|x \in D\}$ such that

$$\sum_{x \in D} T(x)p(x) = \bar{T}.$$

This problem is ill-posed, for in general the number of unknowns ($p(x)$'s), namely $\#D$, is much larger than the number of equations (moments), $L + 1$.²

To circumvent this difficulty, we apply Jaynes (1957)'s maximum entropy principle (MaxEnt). Jaynes proposed that when we want to assign probabilities $\mathbf{p} = (p_1, \dots, p_N)$, given some ‘background information’ (such as moment constraints), we should

choose the least informative distribution by maximizing the Shannon (1948) entropy

$$H(\mathbf{p}) = - \sum_{n=1}^N p_n \log p_n. \quad (4)$$

MaxEnt has been subsequently generalized and axiomatized in such a way to minimize the Kullback–Leibler information of $\mathbf{p} = (p_1, \dots, p_N)$ given the prior distribution³ $\mathbf{q} = (q_1, \dots, q_N)$,

$$H(\mathbf{p}|\mathbf{q}) = \sum_{n=1}^N p_n \log \frac{p_n}{q_n}, \quad (5)$$

subject to the constraints specified in the underlying problem (Caticha and Giffin, 2006). The Shannon entropy (4) can be interpreted as the (negative of) Kullback–Leibler information (5) corresponding to the uniform prior $\mathbf{q} = (1/N, \dots, 1/N)$ modulo an additive constant. MaxEnt methods can be justified in a number of ways⁴ and has been successfully applied in many fields including economics and finance.⁵

The problem can now be formalized as follows. The researcher is given a discrete set $D \subset \mathbb{R}^K$, a prior distribution $\{q(x)|x \in D\}$, a function determining the moments $T: \mathbb{R}^K \rightarrow \mathbb{R}^L$, and moments $\bar{T} \in \mathbb{R}^L$. The objective is to obtain the least informative posterior (in the sense of the Kullback–Leibler information) $\{p(x)|x \in D\}$ that matches the prescribed moments, that is,

$$\begin{aligned} \min_{\{p(x)\}} \sum_{x \in D} p(x) \log \frac{p(x)}{q(x)} \\ \text{subject to } \sum_{x \in D} T(x)p(x) = \bar{T}, \sum_{x \in D} p(x) = 1, p(x) \geq 0. \end{aligned} \quad (P)$$

Returning to the original problem of approximating a density f , if the true density is totally unknown, there is no reason to discriminate one point $x \in D$ over another, so it is natural to choose the uniform prior $q(x) = 1/N$. If the true density f is known (either exogenously given in the model or nonparametrically estimated from data), we can take D to be a lattice on $\mathbb{R}^K$ and take the prior $q(x) = f(x)/\sum_{x \in D} f(x)$, proportional to the given density. If the researcher already has an integration (quadrature) formula

$$\int g(x)f(x)dx \approx \sum_{x \in D} w(x)g(x)f(x), \quad (6)$$

where $\{w(x)\}_{x \in D}$ are weights and g is the integrand (here we regard g as a function whose expectation with respect to the density f we want to compute), then one can use $q(x) = w(x)f(x)$. Note that the “proportional” case is a special case by letting the weighting function $w(x)$ be the constant $1/\sum_{x \in D} f(x)$.

As is common in maximum entropy and Bayesian inference, we do not address how to choose the prior $q(x)$ but take the prior $q(x)$ as given. Instead we fine-tune the probabilities $q(x)$ and obtain the optimal (least informative) posterior $p(x)$ by solving the problem (P).

2.2. Solution

The following theorem shows how to solve problem (P).

³ The maximum entropy literature typically refers to the starting distribution $\mathbf{q}$ as “prior”. This usage (although it may appear unfamiliar) does not contradict with that of Bayesian inference, for Caticha and Giffin (2006) proved that Bayes's rule is implied by minimizing the K–L information.

⁴ In this paper we use MaxEnt merely as a tool. Interested readers can refer to Jaynes (2003) for interpretations, Van Campenhout and Cover (1981) for the relation to Bayesian inference, and Shore and Johnson (1980), Caticha and Giffin (2006), and Knuth and Skilling (2010) for axiomatic approaches.

⁵ See Shore and Johnson (1980) for a short review of applications and Buchen and Kelly (1996), Wu (2003), Veldkamp (2011), and Cabrales et al. (forthcoming) for applications in economics.

² The “+1” comes from accounting the probabilities $\sum_{x \in D} p(x) = 1$.

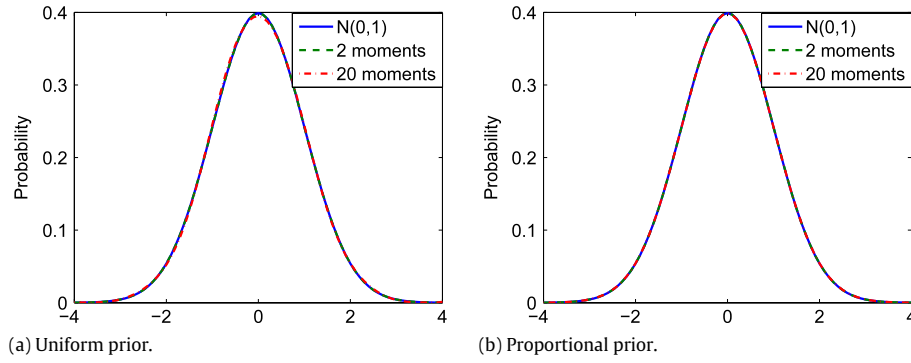


Fig. 1. Implementation of the minimum Kullback–Leibler information algorithm for the standard normal distribution. Blue solid: true distribution, green dashed: moments of order up to 2, red dash-dot: moments of order up to 20.

Theorem 1. Suppose that problem (P) has a solution.⁶ Then the solution is given by

$$p(x) = \frac{q(x)e^{\bar{\lambda}'T(x)}}{\sum_{x \in D} q(x)e^{\bar{\lambda}'T(x)}},$$

where

$$\bar{\lambda} = \arg \min_{\lambda \in \mathbb{R}^L} \left[-\lambda' \bar{T} + \log \left(\sum_{x \in D} q(x) e^{\lambda' T(x)} \right) \right]. \quad (D)$$

Proof. See Appendix. $\square$

Remark. Since in the above solution $p(x)$ is homogeneous in $\{q(x)\}$ and we do not use the condition $\sum_{x \in D} q(x) = 1$ in the proof, it is unnecessary to normalize the prior $\{q(x)\}$.

Note the simplicity of the dual problem (D) compared to the primal problem (P): while the latter is a *constrained* optimization problem with typically a large number of unknowns ($\#D$), the former is an *unconstrained* optimization problem with typically a small number of unknowns (L).

Theorem 1 assumes that the primal problem (P) has a solution, which is not obvious at all when the moment defining function $T(x)$ is complicated. The following theorem directly shows the existence and uniqueness of the solution to the dual problem, and therefore we need not worry about the feasibility of the primal problem.

Theorem 2. Let $C \subset \mathbb{R}^L$ be the convex hull of $T(D)$, i.e., the set consisting of all points of the form $y = \sum_{x \in D} \alpha(x) T(x)$, where $\alpha(x) \geq 0$ and $\sum_{x \in D} \alpha(x) = 1$. If $\bar{T}$ is an interior point of C , then

1. the objective function in (D) is continuous and strictly convex,
2. $\bar{\lambda}$ exists and is unique.

Proof. See Appendix. $\square$

2.3. Numerical implementation

The following algorithm gives the discrete approximation of a probability distribution with density f . We consider three cases depending of whether f is known or unknown and one already has a quadrature formula or not, but in any case the function specifying the moments $T: \mathbb{R}^K \rightarrow \mathbb{R}^L$ and the moments $\bar{T}$ are given.

Step 1. (a) (Uniform case) If the density f is unknown, take any discrete set $D \subset \mathbb{R}^K$ and set the uniform prior $q(x) = 1$ ($x \in D$).

(b) (Proportional case) If the density f is known, take a lattice $D \subset \mathbb{R}^K$ and set the prior proportional to the density, $q(x) = f(x)$ ($x \in D$).

(c) If one already has a quadrature formula (6), set $q(x) = w(x)f(x)$.

Step 2. Solve for

$$\bar{\lambda} = \arg \min_{\lambda \in \mathbb{R}^L} \left[-\lambda' \bar{T} + \log \left(\sum_{x \in D} q(x) e^{\lambda' T(x)} \right) \right]$$

using any optimization routine. Since this is an unconstrained convex optimization problem, global convergence is guaranteed by algorithms such as the Newton–Raphson or the steepest descent methods.

Step 3. The approximating probabilities $\{p(x)|x \in D\}$ are given by

$$p(x) = \frac{q(x)e^{\bar{\lambda}'T(x)}}{\sum_{x \in D} q(x)e^{\bar{\lambda}'T(x)}}.$$

3. Examples

As an application, we approximate the standard normal distribution, the beta distribution, and an AR(1) process. We consider two cases, with density known (proportional prior) or unknown (uniform prior).

3.1. Normal and beta distributions

For the standard normal distribution we choose equally spaced points on $[-4, 4]$ with distance $d = 0.1$ using moments of order up to 20. Fig. 1(a) and (b) show the true density and the probability functions (multiplied by $1/d = 10$ to match the scale) corresponding to the uniform and the proportional priors. In either case the true density and the approximations are virtually indistinguishable.

For the beta distribution, we approximate $B(2, 4)$ using equally spaced points on $[0, 1]$ with distance $d = 0.02$. With the uniform prior, the approximation gets better as the number of moments used increases in Fig. 2(a), but even with 20 moments the difference is still visible. With the proportional prior, the true density and the approximations are virtually indistinguishable in Fig. 2(b).

The reason why the approximation is good in either case for the normal distribution is because the normal distribution is the maximum entropy density given the first and second moments, so the approximations using the uniform and proportional prior coincide if the moments of at least up to the second order are specified. On the other hand, since the beta distribution is not a maximum entropy density (with polynomial moments), the approximating probabilities are dependent on the choice of the prior, which is why using the proportional prior (which uses prior information) gives a better approximation.

⁶ Since the objective function in (P) is continuous, strictly convex, and the constraints are linear, (P) has a unique solution provided that the constraint set is nonempty.

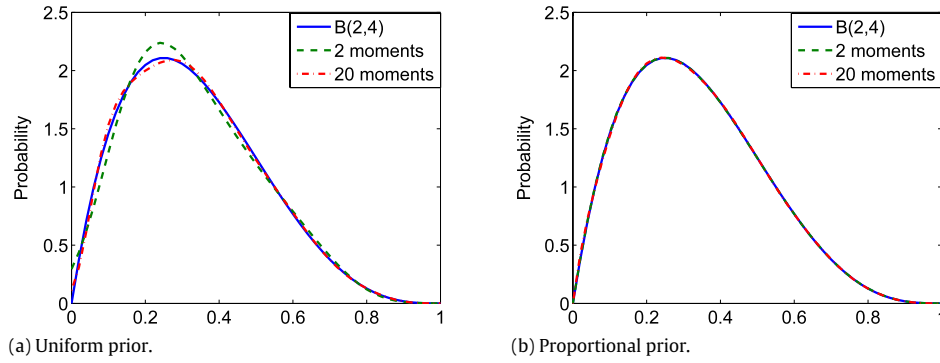


Fig. 2. Implementation of the minimum Kullback-Leibler information algorithm for the beta distribution with parameter $(\alpha, \beta) = (2, 4)$. Blue solid: true distribution, green dashed: moments of order up to 2, red dash-dot: moments of order up to 20.

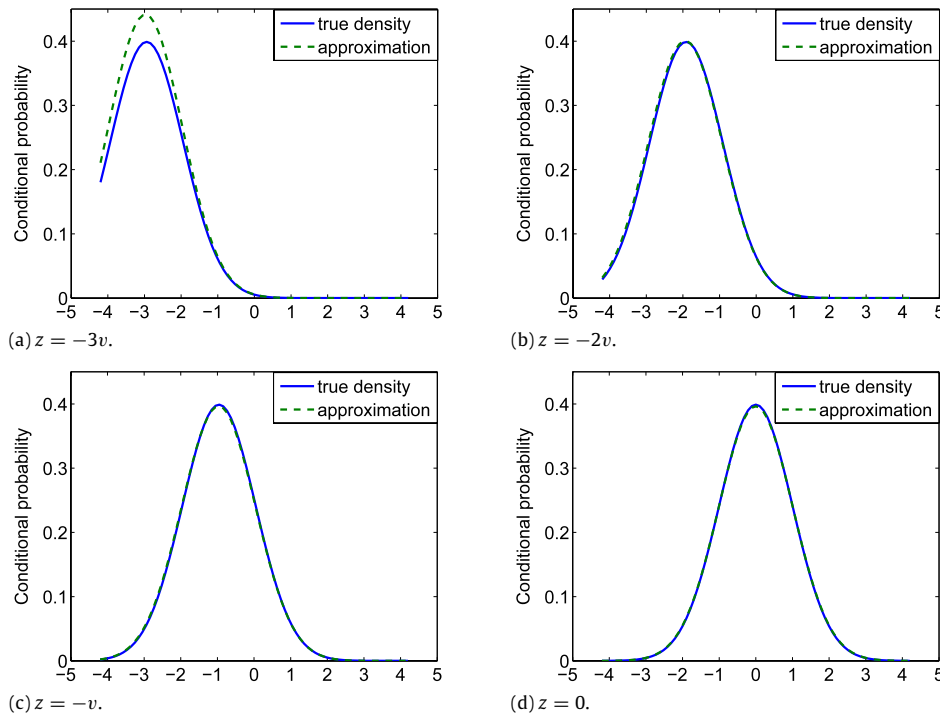


Fig. 3. Distribution of $z' = \rho z + \epsilon$ conditional on $z = -3v, -2v, -v, 0$. Blue solid: true density, green dashed: approximation using the first and second moments.

More generally, consider a canonical probability density function of the exponential family

$$f(x|\theta) = \frac{h(x)e^{\theta'T(x)}}{\int h(x)e^{\theta'T(x)}dx},$$

where $\theta \in \mathbb{R}^L$ is a parameter, $x \in \mathbb{R}^K$, and $T : \mathbb{R}^K \rightarrow \mathbb{R}^L$. By the continuous analogue of Theorem 1, f is the solution to

$$\min_g \int g \log \frac{g}{h} \quad \text{subject to} \quad \int Tg = \bar{T}, \quad \int g = 1, g \geq 0$$

for $\bar{T} = \frac{\partial}{\partial \theta} \log \left(\int h(x)e^{\theta'T(x)}dx \right)$. Therefore $f(x|\theta)$ can be interpreted as the minimum K-L information distribution with “prior” $h(x)$ and the moment defining function $T(x)$. If we use the proportional prior $q(x) = h(x)$ and the same moment $T(x)$ in the discrete approximation, the resulting probability function is of the form

$$p(x) = \frac{h(x)e^{\lambda'T(x)}}{\sum_{x \in D} h(x)e^{\lambda'T(x)}},$$

which has the same functional form as $f(x|\theta)$ except the difference between the integral and the summation. Therefore we can expect

that our method gives an excellent approximation to distributions of the exponential family when we use the relevant moments. (For the Gaussian distribution the relevant moment is $T(x) = (x, x^2)$ and for the beta distribution it is $T(x) = (\log x, \log(1-x))$.)

3.2. AR(1) process

For the AR(1) process $z' = \rho z + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$, we compute the approximating probabilities as follows. Since the joint distribution of (z, z') is $N(0, \Sigma)$ with variance-covariance matrix $\Sigma = \frac{\sigma^2}{1-\rho^2} \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$, we approximate the joint distribution using the first and the second moments ($2(2+3)/2 = 5$ moments in total) and the uniform prior.⁷ We set $\rho = 0.7$, $\sigma = 1$, and choose the range of 4 standard deviation of the stationary distribution (which is $N(0, v^2)$, where the standard deviation is $v = \sigma/\sqrt{1-\rho^2}$) with 81 equally spaced points. Fig. 3 plots the distribution of $z' = \rho z + \epsilon$

⁷ We omit the case with the proportional prior because, by the same reason as discussed when approximating $N(0, 1)$, the results corresponding to the uniform and the proportional priors coincide when approximating a normal distribution if the moments of at least up to the second order are specified.

Table 1
Comparison of our method to existing method.

	Our method	Existing method
Dimension	Arbitrary	Arbitrary
# of moments	Arbitrary	Arbitrary
Distribution	Arbitrary	Arbitrary
Set of moments	Arbitrary	Polynomial
Quadrature formula	Arbitrary	Gaussian quadrature
Optimization problem	<i>Unconstrained (D)</i>	Linear and nonnegativity constraints
Solution	<i>Uniquely exists</i>	Existence and uniqueness unclear

conditional on $z = -3v, -2v, -v, 0$. The approximation is very good for $z \in [-2v, 2v]$ which corresponds to the 95% interval.

The method can be generalized to VAR processes in an obvious way.

4. Concluding remarks

This paper has proposed a numerical method to approximate a given distribution by a discrete distribution with prescribed moments by minimizing the Kullback–Leibler information relative to the prior distribution. The advantages of our method are that it works on any discrete set of any dimension with any moments and is computationally simple. Table 1 compares our method with the existing methods (such as DeVuyst and Preckel, 2007), where we emphasize the advantage of our method in italics.

In this paper we have interpreted “approximation” by matching prescribed moments, but what we are really interested in is the approximation of the probability itself. Numerical examples suggest that the approximation of probabilities are accurate, at least with the proportional prior. We also have some preliminary results on the convergence of the approximating distribution to the given distribution and on the theoretical error bounds, but presenting them are beyond the scope of this paper.

Acknowledgments

We thank Mohsen Bouaissa and the editor for suggestions that improved the paper. This work was supported by the Nakajima Foundation and JSPS KAKENHI Grant Number 24760064.

Appendix

Proof of Theorem 1. To solve (P), we use Fenchel duality. Letting

$$h(p) = \begin{cases} p \log \frac{p}{q}, & (p \geq 0) \\ \infty, & (p < 0) \end{cases}$$

we can ignore the constraint $p(x) \geq 0$ in (P). Since the convex conjugate function of h is

$$h^*(y) = \sup_{p \in \mathbb{R}} [py - h(p)] = \sup_{p \geq 0} \left[py - p \log \frac{p}{q} \right] = qe^{y-1},$$

the dual problem of the primal problem (P) is given by⁸

$$\max_{\lambda \in \mathbb{R}^L, \mu \in \mathbb{R}} \left[\mu + \lambda' \bar{T} - \sum_{x \in D} q(x) e^{\mu + \lambda' T(x) - 1} \right], \quad (D')$$

where λ, μ are Lagrange multipliers on the moment constraint $\sum_{x \in D} T(x)p(x) = \bar{T}$ and the constraint for accounting the probability $\sum_{x \in D} p(x) = 1$. Since the primal problem (P) is a finite-dimensional convex optimization problem with linear constraints,

the regularity condition of the Fenchel duality theorem holds whenever the constraint set is nonempty. Assuming this is the case, the dual problem (D') has a solution (because the constraint set of the primal problem (P) is nonempty and compact) and the primal and dual values coincide. Since (λ, μ) are unconstrained, the maximum is obtained by differentiating and setting the derivative equal to zero. Partially differentiating (D') with respect to μ , we obtain

$$1 - \sum_{x \in D} q(x) e^{\mu + \lambda' T(x) - 1} = 0$$

$$\iff \mu = 1 - \log \left(\sum_{x \in D} q(x) e^{\lambda' T(x)} \right).$$

Substituting this into (D'), after some algebra we obtain

$$H_{\min} := \min \left\{ \sum_{x \in D} p(x) \log \frac{p(x)}{q(x)} \mid \sum_{x \in D} T(x)p(x) = \bar{T}, \right. \\ \left. \sum_{x \in D} p(x) = 1, p(x) \geq 0 \right\}$$

$$= \max_{\lambda \in \mathbb{R}^L} \left[\lambda' \bar{T} - \log \left(\sum_{x \in D} q(x) e^{\lambda' T(x)} \right) \right], \quad (7)$$

which is equivalent to (D). Differentiating (7) with respect to λ and setting the derivative equal to zero, we obtain

$$\sum_{x \in D} T(x) \frac{q(x) e^{\lambda' T(x)}}{\sum_{x \in D} q(x) e^{\lambda' T(x)}} = \bar{T}. \quad (8)$$

Setting $p(x) = q(x) e^{\lambda' T(x)} / \sum_{x \in D} q(x) e^{\lambda' T(x)}$, (8) implies that $\sum_{x \in D} T(x)p(x) = \bar{T}$. Furthermore, direct substitution shows that $\sum_{x \in D} p(x) = 1$ and

$$\sum_{x \in D} p(x) \log \frac{p(x)}{q(x)} = \lambda' \bar{T} - \log \left(\sum_{x \in D} q(x) e^{\lambda' T(x)} \right),$$

so $\{p(x)\}$ attains the minimum in (7). Therefore $\{p(x)\}$ is the (unique) solution to (P). $\square$

Proof of Theorem 2. Let Q be the objective function in (D). The continuity of Q is trivial. Since C has an interior point, it has full dimension. Hence by Proposition B.4 in Toda (2010), $Q(\lambda)$ is strictly convex. Therefore the minimum of $Q(\lambda)$ (if it exists) is unique. To show the existence of a minimum, it suffices to show that $Q(\lambda) \rightarrow \infty$ as $\|\lambda\| \rightarrow \infty$.

Since $\bar{T}$ is an interior point of C , we can take $\epsilon > 0$ such that $\bar{T} + \epsilon \xi$ belongs to the interior of C whenever $\xi \in \mathbb{R}^L$ and $\|\xi\| = 1$ (Euclidean norm). By the definition of the convex hull, for such ξ we can take weights $\{\alpha(x)\}$ (dependent of ξ) such that

$$\bar{T} + \epsilon \xi = \sum_{x \in D} \alpha(x) T(x).$$

For any $x \in D$, by dropping all points except x in the definition of Q , we obtain

$$Q(\lambda) = -\lambda' \bar{T} + \log \left(\sum_{x \in D} q(x) e^{\lambda' T(x)} \right)$$

$$\geq -\lambda' \bar{T} + \lambda' T(x) + \log q(x).$$

Let $q_{\min} = \min_{x \in D} q(x) > 0$. Multiplying both sides by $\alpha(x)$ and summing over $x \in D$, we obtain

$$Q(\lambda) \geq \lambda' \left(\sum_{x \in D} \alpha(x) T(x) - \bar{T} \right) + \log q_{\min}$$

$$= \epsilon \lambda' \xi + \log q_{\min}.$$

⁸ See Corollary 2.6 of Borwein and Lewis (1991).

Since ξ is arbitrary, letting $\xi = \lambda / \|\lambda\|$, we obtain

$$Q(\lambda) \geq \epsilon \|\lambda\| + \log q_{\min}.$$

Hence in minimizing Q over $\mathbb{R}^L$, we may restrict attention to a compact set. Combined with the strict convexity and continuity of Q , $Q(\lambda)$ attains a unique minimum over $\lambda \in \mathbb{R}^L$. $\square$

References

- Adda, Jérôme, Cooper, Russel W., 2003. *Dynamic Economics: Quantitative Methods and Applications*. MIT Press, Cambridge, MA.
- Borwein, Jonathan M., Lewis, Adrian S., 1991. Duality relationships for entropy-like minimization problems. *SIAM Journal on Control and Optimization* 29 (2), 325–338. <http://dx.doi.org/10.1137/0329017>.
- Buchen, Peter W., Kelly, Michael, 1996. The maximum entropy distribution of an asset inferred from option prices. *Journal of Financial and Quantitative Analysis* 31 (1), 143–159. <http://dx.doi.org/10.2307/2331391>.
- Cabrales, Antonio, Gossner, Olivier, Serrano, Roberto, Entropy and the value of information for investors. *American Economic Review* (forthcoming).
- Caticha, Ariel, Giffin, Adom, 2006. Updating probabilities. In: Mohammad-Djafari, Ali (Ed.), *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*. In: AIP Conference Proceedings, vol. 872. pp. 31–42. <http://dx.doi.org/10.1063/1.2423258>.
- DeVuyst, Eric A., Preckel, Paul V., 2007. Gaussian cubature: a practitioner's guide. *Mathematical and Computer Modelling* 45 (7–8), 787–794. <http://dx.doi.org/10.1016/j.mcm.2006.07.021>.
- Jaynes, Edwin T., 1957. Information theory and statistical mechanics. *Physical Review* 106 (4), 620–630. <http://dx.doi.org/10.1103/PhysRev.106.620>.
- Jaynes, Edwin T., 2003. In: Larry Bretthorst, G. (Ed.), *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge, UK.
- Knuth, Kevin H., Skilling, John, 2010. Foundations of inference. URL <http://arxiv.org/abs/1008.4831v1>.
- Miller III, Allen C., Rice, Thomas R., 1983. Discrete approximations of probability distributions. *Management Science* 29 (3), 352–362.
- Shannon, Claude E., 1948. A mathematical theory of communication. *Bell System Technical Journal* 27, 379–423. 623–656.
- Shore, John E., Johnson, Rodney W., 1980. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on Information Theory* IT-26 (1), 26–37.
- Tauchen, George, 1986. Finite state Markov-chain approximations to univariate and vector autoregressions. *Economics Letters* 20 (2), 177–181. [http://dx.doi.org/10.1016/0165-1765\(86\)90168-0](http://dx.doi.org/10.1016/0165-1765(86)90168-0).
- Toda, Alexis Akira, 2010. Existence of a statistical equilibrium for an economy with endogenous offer sets. *Economic Theory* 45 (3), 379–415. <http://dx.doi.org/10.1007/s00199-009-0493-6>.
- Van Campenhout, Jan M., Cover, Thomas M., 1981. Maximum entropy and conditional probability. *IEEE Transactions on Information Theory* IT-27 (4), 483–489.
- Veldkamp, Laura L., 2011. *Information Choice in Macroeconomics and Finance*. Princeton University Press, Princeton, NJ.
- Wu, Ximing, 2003. Calculation of maximum entropy densities with application to income distribution. *Journal of Econometrics* 115 (2), 347–354. [http://dx.doi.org/10.1016/S0304-4076\(03\)00114-3](http://dx.doi.org/10.1016/S0304-4076(03)00114-3).