# Iterative approximation of statistical distributions and relation to information geometry

C.T.J. Dodson*, H. Wang**
*Department of Mathematics,
**Department of Paper Science,
University of Manchester Institute of Science and Technology,
Manchester M60 1QD, UK dodson@umist.ac.uk

*February 11, 2000*

**Abstract**

*The optimal control of stochastic processes through sensor estimation of probability density functions is given a geometric setting via information theory and the information metric. Information theory identifies the exponential distribution as the maximum entropy distribution if only the mean is known and the gamma distribution if also the mean logarithm is known. The surface representing gamma models has a natural Riemannian information metric. The exponential distributions form a 1-dimensional subspace of the 2-dimensional space of all gamma distributions, so we have an isometric embedding of the random model as a subspace of the gamma models. This geometry provides an appropriate structure on which to represent the dynamics of a process and algorithms to control it.*

*This short paper presents a comparative study the parameter estimation performance between the goedesic equation and the B-spline function approximations when they are used to optimize the parameters of the gamma family distributions. In this case, the B-spline functions are first used to approximate the gamma probability density function on a fixed length interval; then the coefficients of the approximation are related, through mean and variance calculations, to the two parameters (i.e., $\mu$ and beta) in gamma distributions. A gradient based parameter tuning method has been used to produce the trajectories for $(\mu, \beta)$ when B-spline functions are used, and desired results have been obtained which are comparable to the trajectories obtained from the geodesic equation*

**Keywords:** INFORMATION GEOMETRY, GAMMA DISTRIBUTION, B-SPLINE APPROXIMATION, PARAMETER ESTIMATION, PROCESS CONTROL

## 1 Introduction

Certain continuous processes involve parametric statistical models to represent dynamic stochastic features of one and two-dimensional time series or textures. In such cases the sampling of the distributed variable may reveal natural non-Gaussian behaviour to optimize. A typical example is when an optical turbidity sensor samples a particle size distribution arising from some comminution process; such distributions are typically of skew type. Indeed, Jaynes [8] provided the foundation for information theoretic methods in, among other things, Bayes hypothesis testing—as described by Tribus et al. [11, 12]. Given a set of observed values $< g_\alpha >$ for functions $g_\alpha$ of the random variable $x$, we seek a 'least prejudiced' probability density $p$ satisfying

$$< g_\alpha >= \int p(x)\, g_\alpha(x)\, dx \quad \text{for } \alpha = 1, 2, \ldots, N \tag{1}$$

Jaynes showed that this occurs if we choose a $p$ that maximizes Shannon's entropy function

$$S = -\int p(x)\log(p(x))\,dx \tag{2}$$

It turns out [12] that if we have no data on observed functions of $x$, (so the set of equations (1) is empty) then the maximum entropy choice gives the exponential distribution. If we have estimates of the first two moments of the distribution of $x$, then we obtain the (truncated) Gaussian. If we have estimates of the mean and mean logarithm of $x$, then the maximum entropy choice is the gamma distribution.

Recent work of Wang [13, 14, 15] has used B-spline bump functions to decompose an arbitrary probability density function and hence optimize its control. Here we consider the information theoretic entropy and geometry of parametric statistical models and offer a Riemannian surface on which dynamics may be studied and optimized. In the sequel we shall consider the particular case of the gamma distribution for several reasons:

- the exponential distributions form a subclass of gamma distributions and exponential distributions represent intervals between events in a Poisson (ie. 'random') process
- lognormal distributions may be well-approximated by gamma distributions
- products of gamma distributions are well-approximated by gamma distributions
- stochastic porous media have been modeled using gamma distributions [5, 6].

Other parametric statistical models based on different distributions may be treated in a similar way.

## 2 Gamma models and their geometry

The family of gamma distributions has event space $\Omega = R^+$, parameters $\mu, \beta \in R^+$ and probability density functions given by

$$f(x;\mu,\beta) = \left(\frac{\beta}{\mu}\right)^{\beta} \frac{x^{\beta-1}}{\Gamma(\beta)}\, e^{-x\beta/\mu} \tag{3}$$

Then $\bar{x} = \mu$ and $Var(x) = \mu^2/\beta$ and we see that $\mu$ controls the mean of the distribution while the spread and shape is controlled by, $1/\beta$, the square of the coefficient of variation.

The usual Riemannian information metric on the parameter space $\mathcal{S} = \{(\mu,\beta) \in \mathcal{R}^+ \times \mathcal{R}^+\}$ is given by

$$ds^2_{\mathcal{S}} = \frac{\beta}{\mu^2}\,d\mu^2 + \left(\psi'(\beta) - \frac{1}{\beta}\right) d\beta^2 \quad \text{for } \mu,\beta \in R^+, \tag{4}$$

where $\psi(\beta) = \frac{\Gamma'(\beta)}{\Gamma(\beta)}$ is the digamma function, the logarithmic derivative of the gamma function. For more details see [9, 2]. The 1-dimensional subspace parametrized by $\beta = 1$ corresponds to the available 'random' processes and the 'length' of any path in $\mathcal{S}$ is given via (4); locally, minimal paths in $\mathcal{S}$ are given by the geodesics defined by (4). The Gaussian curvature of the surface $\mathcal{S}$ is

$$K(\mu,\beta) = \frac{3 - 4\,\beta\,\psi'(\beta) - \beta^2\,\psi''(\beta)}{2\,\beta\,(-1+\beta\,\psi'(\beta))^2} \quad \text{for } \mu,\beta \in \mathbb{R}^+ \tag{5}$$

$$K_{\mathcal{S}}(\mu,\beta) \;\to\; -1 \text{ as } \beta \to 0 \tag{6}$$

$$K_{\mathcal{S}}(\mu,\beta) \;\to\; -2 \text{ as } \beta \to \infty \tag{7}$$

Further study of the geometry of this parameter space will be reported elsewhere.

## 3 B-spline approximation to $\Gamma$ distributions

The management and performance optimization of a stochastic process represented by gamma parametric models may then be represented through this Riemannian geometry. Other distributions may be treated similarly. In the case of simultaneous handling of two probability density functions, considered by Wang [15], we have a product of two Riemannian spaces and correlations between the two distributions would appear as a twisting of this product.

In this case the family of $\Gamma-$distribution can still be considered. However, since

$$\lim_{x \to +\infty} f(x, \mu, \beta) = 0 \tag{8}$$

for any arbitrarily small $\epsilon > 0$ there is a $b(\epsilon, \mu, \beta) > 0$ such that the following inequality holds

$$|f(x, \mu, \beta)| \leq \epsilon \ \forall x > b(\epsilon, \mu, \beta) \tag{9}$$

This means that we can use the following B-spline functions to approximate the probability density function to give

$$\left| f(x, \mu, \beta) - \sum_{i=1}^{n} w_i B_i(x) \right| \leq \delta \tag{10}$$

where $B_i(x)$ are the pre-specified basis functions defined on $[0, b]$, $w_i$ are the weights (i.e., expansion coefficients) to be trained and $\delta$, which is generally larger than $\epsilon$, is a pre-specified arbitrarily small positive number. Indeed, such an approximation will be used here to represent the coupled links between the two parameters $(\mu, \beta)$ and the distribution density function $f(x, \mu, \beta)$. Since the $\Gamma$ probability density function has an exponential element, in the B-spline approximation a set of truncated Guassian functions of the following form

$$B_i(x) = e^{-\frac{(x-\xi_i)^2}{2\sigma_i^2}}; \ \ \xi_i = \frac{1}{2}(\lambda_i + \lambda_{i+1}) \tag{11}$$

are used, where $\sigma_i$ is a pre-selected small positive number so that the truncated Guassian functions cover the width of the sub-interval $[\lambda_i, \lambda_{i+1}] \in [0, b]$.

Since all the basis functions are pre-specified, different values of $\mu$ and $\beta$ will generate different sets of weights. This means that the weights $w_i$ are functions of the parameter pair $(\mu, \beta)$. As such, the approximation (11) should be further represented as

$$f(x, \mu, \beta) = \sum_{i=1}^{n} w_i(\mu, \beta) B_i(x) + e \tag{12}$$

where $e$ represents the approximation error, which satisfies $|e| \leq \delta$. Figure 1 shows the principle of this approximation when $u(t) \in R^1$. To simplify the presentation, in the rest of the paper the term $e$ in equation (13) will be neglected.

Since

$$\int_0^b f(x, \mu, \beta) dx = 1 \tag{13}$$

holds, this, together with equation (13), means that only $n - 1$ weights are independent. As a result, a reformulation is required to express the expansion (13) by only using the $n - 1$ weights. Indeed, this has been formulated by Wang (1997, 1998). For the completeness of the paper, the result of this formulation is included in the following.

Define

$$b_i = \int_0^b B_i(x) dx \quad (i = 1, 2, ..., n) \tag{14}$$

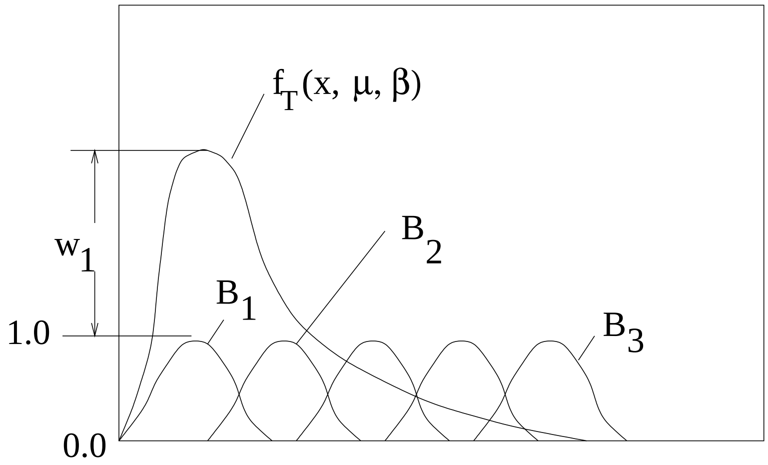


Figure 1: B-splines expression for the truncated probability density function $f_b(x, \mu, \beta)$.

and

$$\begin{aligned}
L(x) &= b_n^{-1} B_n(x) \in R^{1\times 1} \\
C(x) &= (B_1(x) - \frac{B_n(x)}{\int_0^b B_n(x)dy}\int_0^b B_1(x)dx, \\
&\quad B_2(x) - \frac{B_n(x)}{\int_0^b B_n(x)dx}\int_0^b B_2(x)dx, \\
&\quad ..., B_{n-1}(x) - \frac{B_n(x)}{\int_0^b B_n(x)dx}\int_0^b B_{n-1}(x)dx) \in R^{1\times(n-1)}
\end{aligned} \tag{15}$$

then it has been shown (Wang, 1997, 1998) that

$$f(x, \mu, \beta) = C(x)V + L(x) \tag{16}$$

where $V = (w_1, w_2, ..., w_{n-1})$ is the parameter vector. From the definition (15)-(16), it can be shown that $\int_0^b C(x)dx = 0$ and $\int_0^b L(x)dx = 1$. As such, the proposed constraint of equality (14) can be natually guaranteed.

# 4 The relationship between $V$ and $(\mu, \beta)$

It can be seen that in our case

$$\mu = \int_0^b xf(x, \mu, \beta)dx = (\int_0^b xC(x)dx)V + \int_0^b xL(x)dx \tag{17}$$

As a result, it can be further obtained that

$$\beta = \frac{\mu^2}{Var(x)} \tag{18}$$

where, using equality (17),

$$\begin{aligned}
Var(x) &= \int_0^b (x-\mu)^2 f(x, \mu, \beta)dx \\
&= (\int_0^b (x-\mu)^2 C(x)dx)V + \int_0^b (x-\mu)^2 L(x)dx
\end{aligned} \tag{19}$$

As a result, it can be seen that the weight vector $V$ is directly related to $\mu$ and $\beta$.

# 5 Control of the distribution

Once the models such as the one given by equation (17) is formulated, the next question arising from many practical situations is to see how the parameters of a distribution can be selected so that the actual distribution is made as close as possible to a given one. This is a control problem and has many applications in paper and board making (Wang, 1998).

In this paper, we shall still concentrate on the $\Gamma$-distribution and study the trajectory of the parameter tuning when the control of the probability density function is required. For this purpose, it is assumed that the initial $\Gamma$-distribution is characterised by $(\mu_0, \beta_0)$ and the desired distribution is presented by $(\mu_f, \beta_f)$. In fact, since an information - optimal trajectory can be obtained using the geodesic equation in the space of gamma models, focus will be made on

- the evaluation of how the weight vector $V$ of model (17) will behave in tuning the initial distribution to the final distribution, and
- the comparison of the equivalent $(\mu, \beta)$-trajectory of model (17) with those trajectories obtained from the geodesic equation.

Since the weight vector $V$ is directly related to the two parameters in the $\Gamma$-distribution, the tuning rule for the weight vector $V$ will be used. This means that an effective trajectory for $V$ should be chosen such that the following performance function

$$J = \frac{1}{2}\int_0^b (C(x)V + L(x) - f(x, \mu_f, \beta_f))^2 dx \tag{20}$$

can be minimized. This leads to the following application of the gradient rule

$$V_k = V_{k-1} - \eta \frac{\partial J}{\partial V}|_{V=V_{k-1}} \tag{21}$$

where $k = 0, 1, 2, \cdots$ represents the sample number and $\eta > 0$ is a pre-specified learning rate. The gradient in (21) can be further formulated to give

$$\frac{\partial J}{\partial V} = \int_0^b (C(x)V + L(x) - f(x, \mu_f, \beta_f))C^T(x)dx \tag{22}$$

Using (17) and (18)- (19), the corresponding values of $\mu_k$ and $\beta_k$ can be obtained from

$$\mu_k = (\int_0^b xC(x)dx)V_k + \int_0^b xL(x)dx \tag{23}$$

$$\beta_k = \frac{\mu_k^2}{Var(x,k)} \tag{24}$$

where

$$\begin{aligned} Var(x,k) &= (\int_0^b (x-\mu)^2 C(x)dx)V_k \\ &+ \int_0^b (x-\mu)^2 L(x)dx \end{aligned} \tag{25}$$

As a result, the following training algorithm can be obtained which produces the trajectory of the $(\mu_k, \beta_k)$ when the performance function (20) is minimised via the updating rule (21).

1. Select the initial weight vector $V_0$ so that
$$C(x)V_0 + L(x) = f(x, \mu_0, \beta_0) \tag{26}$$
2. Use the training algorithm (21) to update the weight vector $V_k$
3. Calculate $\mu_k$ and $\beta_k$ using (23), (24) - (25)

## 6 Simulation results

### 6.1 The learning results for model (24)

Simulations have been performed in order to illustrate the use of the algorithm in section 7. The initial Γ-distribution is defined by $\mu = 2$ and $\beta = 7$ whilst the final distribution is charaterised by $\mu = 0.5$ and $\beta = 2$, both of these two distributions are shown in Fig. 2. For these two distributions, the lenght of the truncated interval is selected as $b = 5.0$. Fifty truncated guassian basis functions are used to model the truncated distribution with the internal knots be eq ualy distributed. This leads to $\sigma_i = 0.25$ in equation (12).

The learning algorithm in equation (22) is used with the learning $\eta = 0.000001$. The result is shown in Fig. 3 and the final trained vector $V_{6000}$ is given in Fig.4. From these figures, it can be seen that a good learning result has been obtained for $x \leq 4$. This can also be seen in Fig. 4, where the magnidutes of the last few components of $V_k$ are too big, thus leading to a large approximation error when $x \in [4, 5]$.

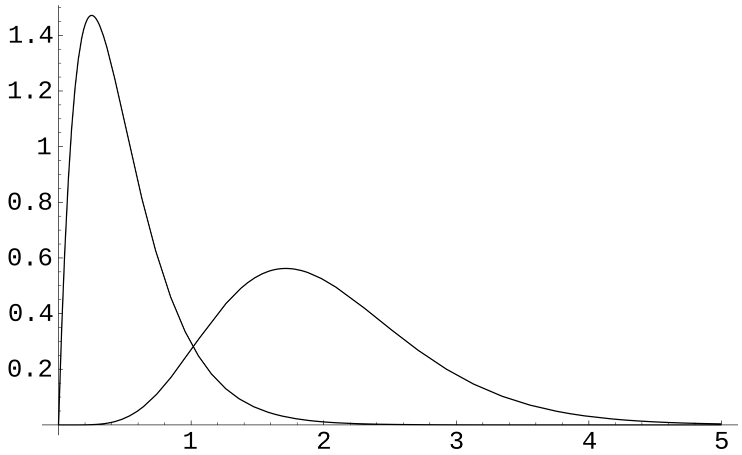


Figure 2: The initial and final truncated probability density function $f_b(x, \mu, \beta)$.

### 6.2 Comparison with the information geometry of parameter space

In Figure 5, the corresponding trajectory of the equivalent $\mu_k$ and $\beta_k$ of subsection 6.1 is shown. From this figure it can be seen that the trajectory is a straight line but does not converge exactly to the final values of $\mu = 0.5$ and $\beta = 2$. However, it can be concluded that the trajectory is linear or almost linear around a neighbourhood of $(\mu, \beta) = (0.5, 2)$. This can be varified by considering the parameter learning algorithm in (22) again. From equation (22), it can be seen that this learning algorithm is linear in parameters $V$. Hence the trajectory is a straight line in the space of $V$. However, this does not necessarily mean that the trajectory is a straight line in $(\mu, \beta)$ space. As such, the trajectory shown in Figure 5 only reflect their local properties around $(\mu, \beta) = (0.5, 2)$.

Instead, in Figure 6 we show the 1-parameter family, or homotopy, of trajectories from the initial to final states given by:

$$\alpha_\epsilon : [0, 1] \to \mathcal{S} : t \mapsto (2 - \frac{3}{2}t, 7 - 5t + \epsilon t(1 - t)), \quad \epsilon \in [-5, 5]. \tag{27}$$

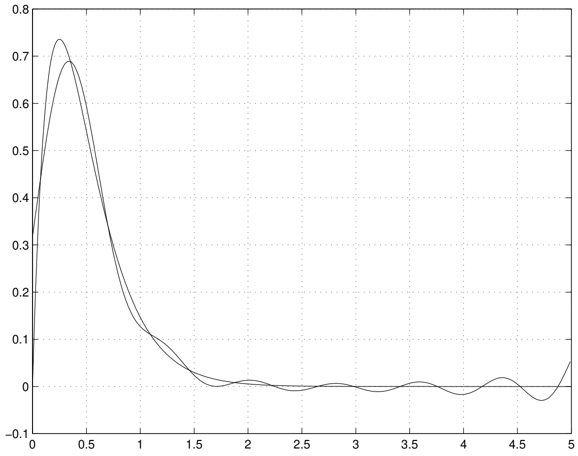


Figure 3: The learning results.

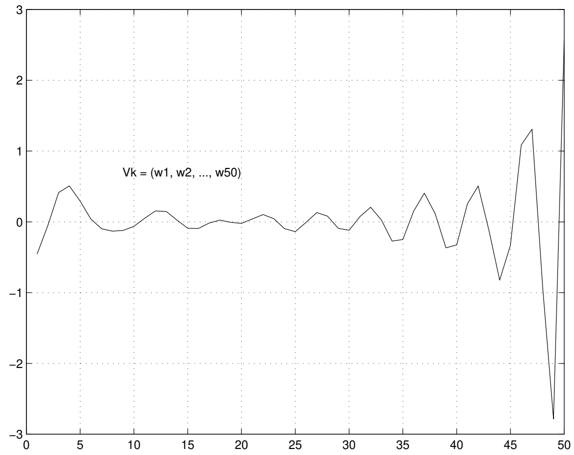


Figure 4: The trained weight vector $V_k$.

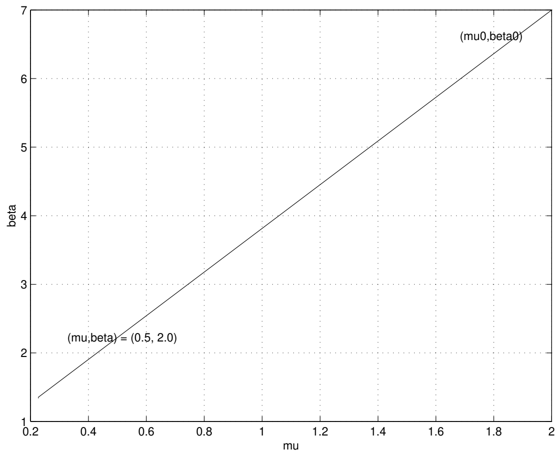


Figure 5: The $\mu_k - \beta_k$ trajectory using the learning algorithm (29).

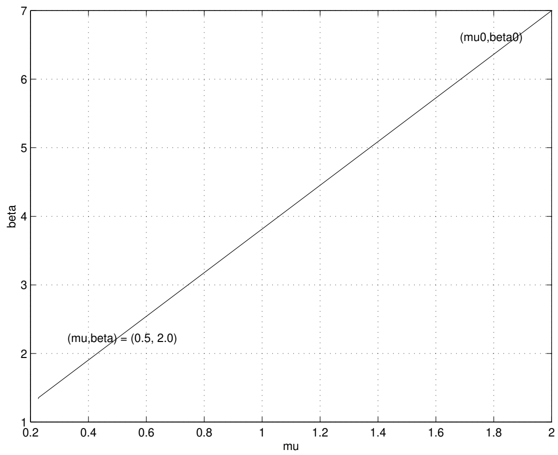


Figure 6: The local $\mu_k - \beta_k$ trajectories obtained from the geodesic equation.

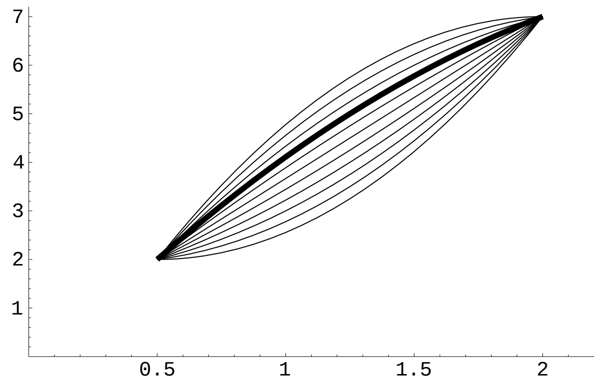

Figure 7: A 1-parameter family of trajectories between the initial and final states; the bold curve has minimal information-length. The simulation trajectory in Figure 8 is the central curve in the family.

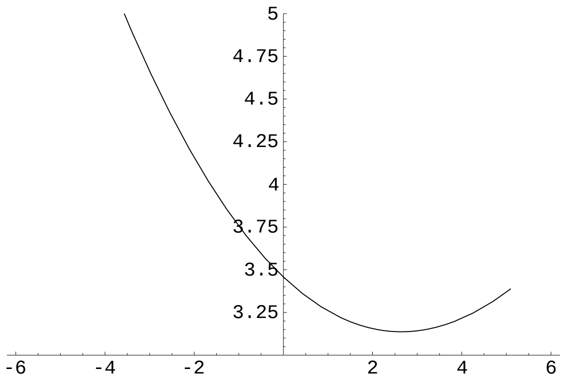

Figure 8: The information arclength $L(\epsilon)$ (up to a monotonic function) for the 1-parameter family of trajectories in Figure 9; the minimal length at $\epsilon \approx 2.5$ corresponds to the bold curve in Figure 9.

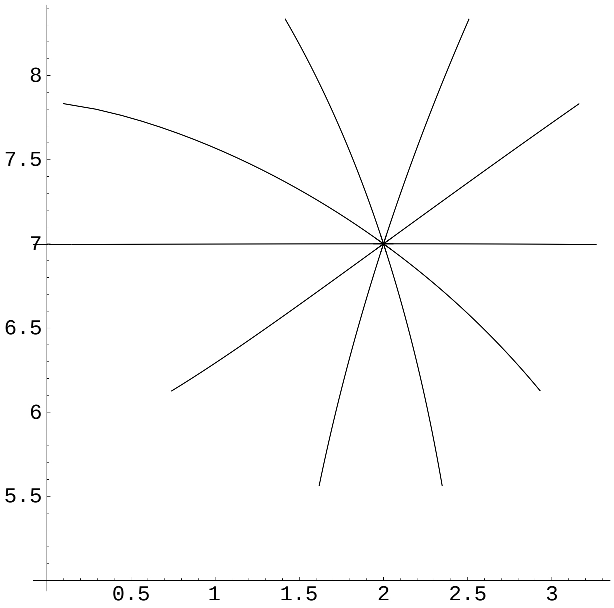


Figure 9: The $\mu_k - \beta_k$ geodesic trajectories near the initial distribution.

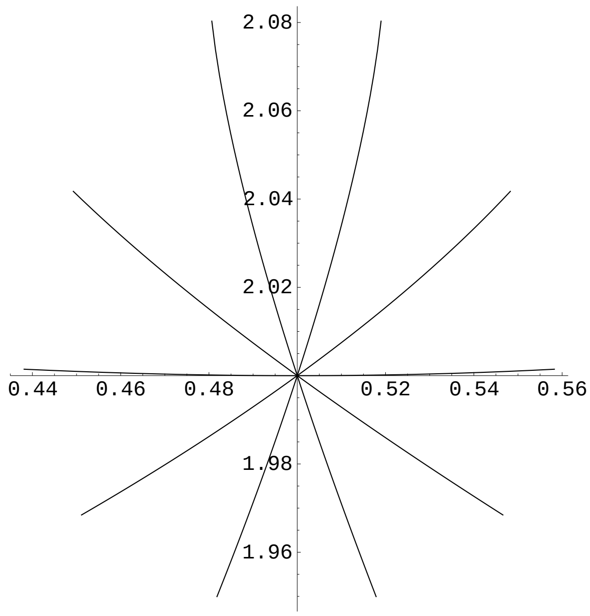


Figure 10: The local $\mu_k - \beta_k$ geodesic trajectories near the final distribution.

This contains the linear simulation trajectory as $\alpha_0$, the central curve. The bold curve in the family (27) of Figure 11 is the trajectory with minimal information-length, close to $\alpha_{2.5}$, as may be seen from Figure 7, which shows a plot of arclength (up to a monotonic function) as a function of the perturbing parameter $\epsilon$.

Figures 8 and 9 show sample sprays of geodesics in the vicinities of the initial and final states. Analytically, the global optimal trajectory will be closer to a piecewise-geodesic curve from $p_0$ to $p_1$ but in general that is hard to find [4].

We see now how to use the information geometry in a practical way to optimise in a maximum likelihood sense the learning trajectory steps:

- Choose initial $p_0$ and final $p_1$ points in the parameter space of probability density functions.
- Create a 1-parameter family of trajectories $\alpha_\epsilon$ from $p_0$ to $p_1$ with $\alpha_0$ the line from $p_0$ to $p_1$.
- Compute the information metric arclength $L(\epsilon)$ of each $\alpha_\epsilon$.
- That $\epsilon = \epsilon*$ having minimum length $L$ is in a maximum likelihood sense a locally optimal trajectory for the process change from $p_0$ to $p_1$.

## References


[1] S-I. Amari. **Differential Geometrical Methods in Statistics** Springer Lecture Notes in Statistics 28, Springer-Verlag, Berlin 1985.

[2] C.T.J. Dodson. Gamma manifolds and stochastic geometry. In: **Proceedings of the Workshop on Recent Topics in Differential Geometry**, Santiago de Compostela 16-19 July 1997. *Public. Depto. Geometría y Topología* 89 (1998) 85-92.

[3] C.T.J. Dodson. Systems of connections for parametric models. In **Proc. Workshop on Geometrization of Statistical Theory** 28-31 October 1987. Ed. C.T.J. Dodson, ULDM Publications, University of Lancaster, 1987, pp 153-170.

[4] C.T.J. Dodson and T. Poston. **Tensor Geometry** Graduate Texts in Mathematics 130, Springer-Verlag, Berlin, Heidelberg, New York 1991.

[5] C.T.J. Dodson and W.W. Sampson. Modeling a class of stochastic porous media. *Appl. Math. Lett.* 10, 2 (1997) 87-89.

[6] C.T.J. Dodson and W.W. Sampson. Spatial statistics of stochastic fibre networks. 1999. *J. Statist. Phys* 96, 1/2 (1999) 447-458.

[7] R.A. Fisher. Theory of statistical estimation. *Proc. Camb. Phil. Soc.* 122 (1925) 700-725.

[8] E.T. Jaynes. Information theory and statistical inference. *The Physical Review* 106 (1957) 620-630 and 108 (1957) 171-190.

[9] S.L. Lauritzen. Statistical Manifolds. In **Differential Geometry in Statistical Inference**, Institute of Mathematical Statistics Lecture Notes, Volume 10, Berkeley 1987, pp 163-218.

[10] C.R. Rao. Information and accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.* 37, (1945) 81-91.

[11] M. Tribus **Thermostatics and Thermodynamics** D. Van Nostrand and Co., Princeton N.J., 1961.)

[12] M. Tribus, R. Evans and G. Crellin. The use of entropy in hypothesis testing. In Proc. **Tenth National Symposium on Reliability and Quality Control** 7-9 January 1964.

[13] H. Wang. Neural network based control for output probability density functions for nonlinear stochastic systems and its applications. In Proc. **6**$^{th}$ **European Congress on Intelligent Techniques and Soft Computing** Aachen, 1998.

[14] H. Wang. Control of the output probability density functions for a class of nonlinear stochastic systems. In Proc. **IFAC Workshop on Algorithms and Architectures for Real Time Control** 1998, pp 113-117.

[15] H. Wang. Detect unexpected changes of particle size distribution in paper-making white water system. In Proc. **IFAC Fault Detection and Diagnosis** Lyon, 1998, pp. 78-84.