Information geometry for testing pseudorandom number generators

C.T.J. Dodson
School of Mathematics, University of Manchester, Manchester M13 9PL, UK
ctdodson@manchester.ac.uk

1 Introduction

The smooth family of gamma probability density functions is given by

$$f : [0,\infty) \to [0,\infty) : x \mapsto \frac{e^{-\frac{x\kappa}{\mu}} x^{\kappa-1} \left(\frac{\kappa}{\mu}\right)^{\kappa}}{\Gamma(\kappa)} \quad \mu,\kappa > 0. \tag{1}$$

Here μ is the mean, and the standard deviation σ, given by $\kappa = (\frac{\mu}{\sigma})^2$, is proportional to the mean. Hence the coefficient of variation $\frac{1}{\sqrt{\kappa}}$ is unity in the case that (1) reduces to the exponential distribution. Thus, $\kappa = 1$ corresponds to an underlying Poisson random process complementary to the exponential distribution. When $\kappa < 1$ the random variable X represents spacings between events that are more clustered than for a Poisson process and when $\kappa > 1$ the spacings X are more uniformly distributed than for Poisson. The case when $\mu = n$ is a positive integer and $\kappa = 2$ gives the Chi-Squared distribution with $n-1$ degrees of freedom; this is the distribution of $\frac{(n-1)s^2}{\sigma_G^2}$ for variances s^2 of samples of size n taken from a Gaussian population with variance σ_G^2.

The gamma distribution has a conveniently tractable information geometry [1, 2], and the Riemannian metric in the 2-dimensional manifold of gamma distributions (1) is

$$[g_{ij}](\mu,\kappa) \quad = \quad = \begin{bmatrix} \frac{\kappa}{\mu^2} & 0 \\ 0 & \frac{d^2}{d\kappa^2}\log(\Gamma) - \frac{1}{\kappa} \end{bmatrix}. \tag{2}$$

So the coordinates (μ,κ) yield an orthogonal basis of tangent vectors, which is useful in calculations because then the arc length function is simply

$$ds^2 = \frac{\kappa}{\mu^2}\, d\gamma^2 + \left(\left(\frac{\Gamma'(\kappa)}{\Gamma(\kappa)} \right)' - \frac{1}{\kappa} \right) d\kappa^2.$$

We note the following important uniqueness property:

Theorem 1.1 (Hwang and Hu [4]) *For independent positive random variables with a common probability density function f, having independence of the sample mean and the sample coefficient of variation is equivalent to f being the gamma distribution.*

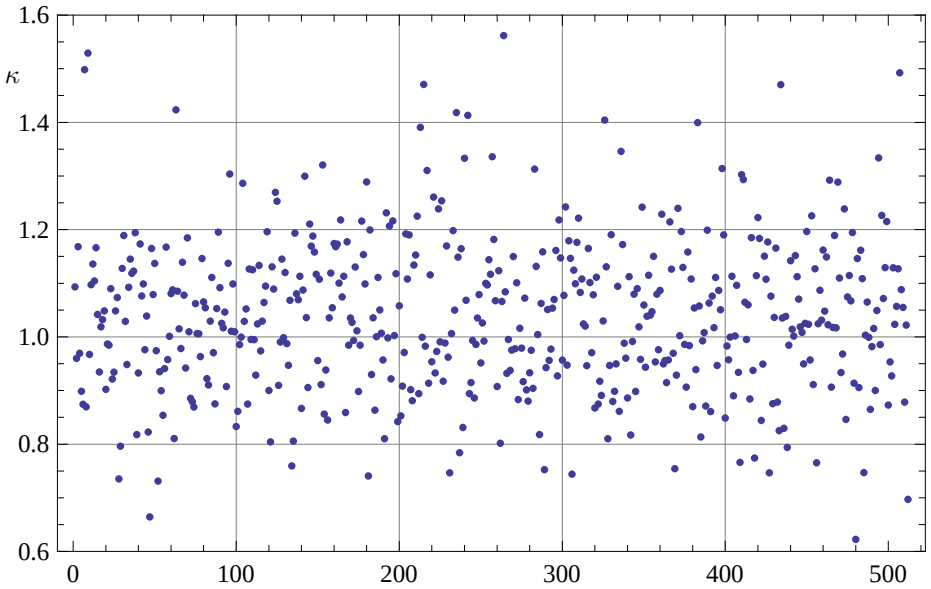

Figure 1: *Maximum likelihood gamma parameter* κ *fitted to separation statistics for simulations of Poisson random sequences of length* 100000 *for an element with expected parameters* $(\mu, \kappa) = (511, 1)$. *These simulations used the pseudo-random number generator in Mathematica [7].*

This property is one of the main reasons for the large number of applications of gamma distributions: many near-random natural processes have standard deviation approximately proportional to the mean [2]. Given a set of identically distributed, independent data values $X_1, X_2, \ldots, X_n$, the 'maximum likelihood' or 'maximum entropy' parameter values $\hat{\mu}, \hat{\kappa}$ for fitting the gamma distribution (1) are computed in terms of the mean and mean logarithm of the X_i by maximizing the likelihood function

$$L_f(\mu, \kappa) = \prod_{i=1}^{n} f(X_i; \mu, \kappa).$$

By taking the logarithm and setting the gradient to zero we obtain

$$\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{3}$$

$$\begin{aligned} \log \hat{\kappa} - \frac{\Gamma'(\hat{\kappa})}{\Gamma(\hat{\kappa})} &= \log \bar{X} - \frac{1}{n} \sum_{i=1}^{n} \log X_i \\ &= \log \bar{X} - \overline{\log X}. \end{aligned} \tag{4}$$

2 Neighbourhoods of randomness in the gamma manifold

In a variety of contexts in cryptology for encoding, decoding or for obscuring procedures, sequences of pseudorandom numbers are generated. Tests for ran-

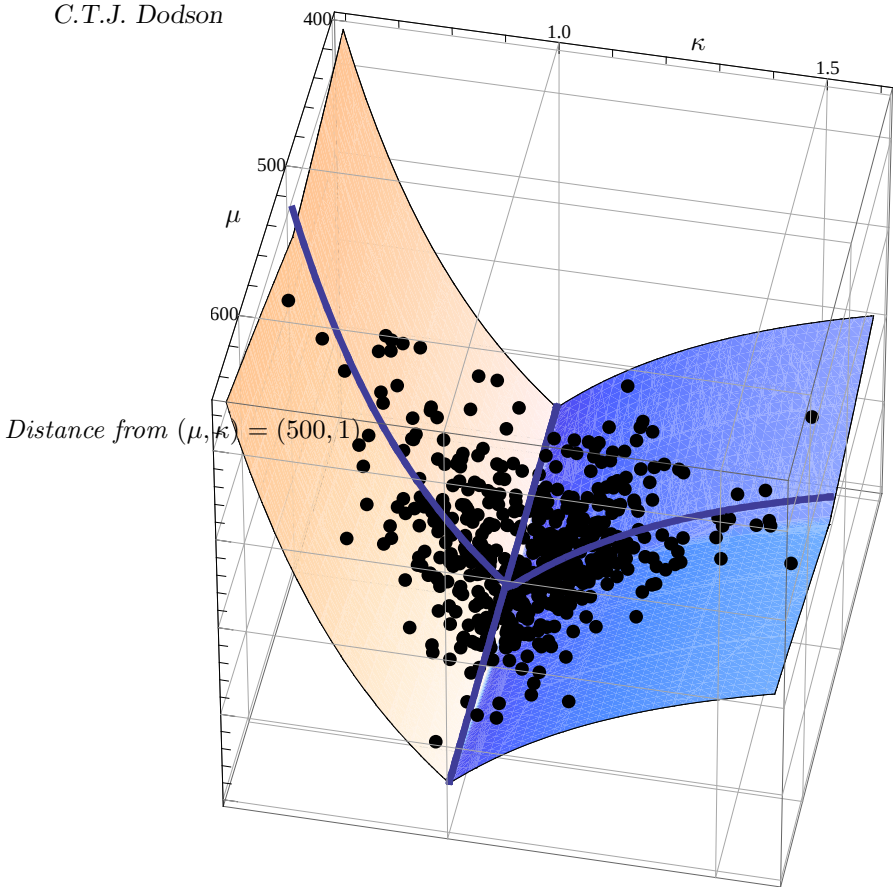

Figure 2: *Distances in the space of gamma models, using a geodesic mesh. The surface height represents upper bounds on distances from* $(\mu, \kappa) = (511, 1)$ *from Equation (5). Also shown are data points from simulations of Poisson random sequences of length* 100000 *for an element with expected separation* $\mu = 511$. *In the limit as the sequence length tends to infinity and the element abundance tends to zero we expect the gamma parameter* κ *to tend to* 1.

domness of such sequences have been studied extensively and the NIST Suite of tests [5] for cryptological purposes is widely employed. Information theoretic methods also are used, for example see Grzegorzewski and Wieczorkowski [3] also Ryabko and Monarev [6] and references therein for recent work. Here we can show how pseudorandom sequences may be tested using information geometry by using distances in the gamma manifold to compare maximum likelihood parameters for separation statistics of sequence elements.

Mathematica [7] simulations were made of Poisson random sequences with length $n = 100000$ and spacing statistics were computed for an element with abundance probability $p = 0.00195$ in the sequence. Figure 1 shows maximum likelihood gamma parameter κ data points from such simulations. In the data from 500 simulations the ranges of maximum likelihood gamma distribution parameters were $419 \leq \mu \leq 643$ and $0.62 \leq \kappa \leq 1.56$.

The surface height in Figure 2 represents upper bounds on information geometric distances from $(\mu, \kappa) = (511, 1)$ in the gamma manifold. This employs

the geodesic mesh function we described in Arwini and Dodson [2].

$$Distance[(511,1),(\mu,\kappa)] \leq \left| \frac{d^2 \log \Gamma}{d\kappa^2}(\kappa) - \frac{d^2 \log \Gamma}{d\kappa^2}(1) \right| + \left| \log \frac{511}{\mu} \right|. \quad (5)$$

Also shown in Figure 2 are data points from the *Mathematica* simulations of Poisson random sequences of length 100000 for an element with expected separation $\gamma = 511$.

In the limit, as the sequence length tends to infinity and the abundance of the element tends to zero, we expect the gamma parameter τ to tend to 1. However, finite sequences must be used in real applications and then provision of a metric structure allows us, for example, to compare real sequence generating procedures against an ideal Poisson random model.

References

[1] S-I. Amari and H. Nagaoka. **Methods of Information Geometry**, American Mathematical Society, Oxford University Press, 2000.

[2] Khadiga Arwini and C.T.J. Dodson. **Information Geometry Near Randomness and Near Independence**. Lecture Notes in Mathematics, Springer-Verlag, New York, Berlin 2008.

[3] P. Grzegorzewski and R. Wieczorkowski. Entropy-based goodness-of-fit test for exponentiality. *Commun. Statist. Theory Meth.* 28, 5 (1999) 1183-1202.

[4] T-Y. Hwang and C-Y. Hu. On a characterization of the gamma distribution: The independence of the sample mean and the sample coefficient of variation. *Annals Inst. Statist. Math.* 51, 4 (1999) 749-753.

[5] A. Rushkin, J. Soto et al. **A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications**. *National Institute of Standards & Technology*, Gaithersburg, MD USA, 2001.

[6] B.Ya. Ryabko and V.A. Monarev. Using information theory approach to randomness testing. Preprint: *arXiv:CS.IT/0504006 v1*, 3 April 2005.

[7] S. Wolfram. **The Mathematica Book** 3^{rd} edition, Cambridge University Press, Cambridge, 1996.