Large spread does not imply Benford's Law

A. Berger

Mathematical and Statistical Sciences
University of Alberta, Edmonton, Canada

Preliminary draft version 11 January 2010

Abstract

Sharp universal bounds are given for the distance between normalised Lebesgue measure on $\mathbb{R}/\mathbb{Z}$ and the distribution of $\log X$ mod 1, where X is uniform. The results dispel the popular belief that a random variable obeys Benford's Law (at least approximately) whenever its spread is large.

For every real number x, the largest integer not larger than x will be denoted by $\lfloor x \rfloor$, and $[\![x]\!] := x - \lfloor x \rfloor$ is the fractional (or non-integer) part of x. The base γ logarithm ($\gamma > 1$) of $x > 0$ is $\log_\gamma x$; if used without a subscript, log symbolises the natural logarithm. The sets of natural, non-negative integer, integer, positive real and real numbers are $\mathbb{N}$, $\mathbb{N}_0$, $\mathbb{Z}$, $\mathbb{R}^+$ and $\mathbb{R}$, respectively.

Given any probability measure μ on $\mathbb{R}$, denote by F_μ its distribution function, that is, $F_\mu(x) = \mu(]-\infty, x])$. For every measurable map $T : \mathbb{R} \to \mathbb{R}$ the probability measure $T\mu$ is defined as $T\mu(B) := \mu\big(T^{-1}(B)\big)$ for all Borel sets B. Specifically, $[\![\mu]\!]$ is, for every μ, concentrated on $[0, 1]$. The uniform distribution on $[a, b]$ with $a < b$ is denoted by $U_{a,b}$. Thus

$$F_{U_{a,b}}(x) = \begin{cases} 0 & \text{if } x < a\,, \\ \dfrac{x-a}{b-a} & \text{if } a \le x < b\,, \\ 1 & \text{if } x \ge b\,. \end{cases}$$

Given any two probability measures μ, ν on $\mathbb{R}$, their Kolmogorov–Smirnov distance $d_\infty(\mu, \nu)$ is

$$d_\infty(\mu, \nu) = \sup_{x \in \mathbb{R}} \big| F_\mu(x) - F_\nu(x) \big|\,;$$

see e.g. [5] for some details on this metric. Recall that μ with $\mu(\mathbb{R}^+) = 1$, or a real random variable with distribution μ, satisfies *Benford's Law* base γ if and only if $\log_\gamma \mu$ is uniform modulo one [2], i.e., if $d_\infty([\![\log_\gamma \mu]\!], U_{0,1})$ equals zero. Contrary to what [4, p.63] may suggest, the uniform distribution $U_{a,b}$ with $a \ge 0$ does not even approximately satisfy Benford's Law for any base γ, no matter how large $b - a$ is.

Theorem 1. *For all $\gamma > 1$ and $0 \le a < b$,*

$$d_\infty\big([\![\log_\gamma U_{a,b}]\!], U_{0,1}\big) \ge C_\gamma > 0\,, \tag{1}$$

where

$$C_\gamma = \frac{1 - \gamma + \log\gamma + (\gamma - 1)\log(\gamma - 1) - (\gamma - 1)\log\log\gamma}{2(\gamma - 1)\log\gamma}\,.$$

E-mail address: aberger@math.ualberta.ca

Proof. To verify (1), assume first that $a = 0$, and let $l := \lfloor \log_\gamma b \rfloor$ and $\delta := [\![\log_\gamma b]\!]$. Thus $b = \gamma^{l+\delta}$, and

$$F_{\log_\gamma U_{0,b}}(x) = \begin{cases} \dfrac{\gamma^x}{b} & \text{if } x < l + \delta\,, \\ 1 & \text{if } x \geq l + \delta\,, \end{cases}$$

from which it follows that, for all $0 \leq x \leq 1$,

$$F_{[\![\log_\gamma U_{0,b}]\!]}(x) = \sum\nolimits_{k\in\mathbb{Z}} \bigl(F_{\log_\gamma U_{0,b}}(x+k) - F_{\log_\gamma U_{0,b}}(k)\bigr) = \begin{cases} \dfrac{\gamma^{1+x-\delta} - \gamma^{1-\delta}}{\gamma - 1} & \text{if } 0 \leq x < \delta\,, \\ 1 - \dfrac{\gamma^{1-\delta} - \gamma^{x-\delta}}{\gamma - 1} & \text{if } \delta \leq x \leq 1\,. \end{cases}$$

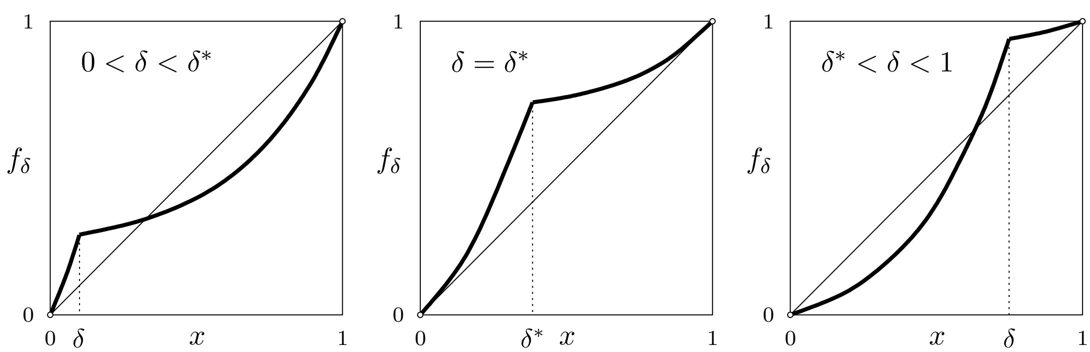
Notice that $F_{[\![\log_\gamma U_{0,b}]\!]}$ does not depend on l. For the sake of brevity, let $f_\delta(x) := F_{[\![\log_\gamma U_{0,b}]\!]}(x)$ and $g_\delta(x) := f_\delta(x) - x$ for all x and δ. Note that f_δ is convex on $[0, \delta]$ and on $[\delta, 1]$, and $f_0(x) = \dfrac{\gamma^x - 1}{\gamma - 1}$, whereas for $0 < \delta < 1$

$$f'_\delta(0+) = f'_\delta(1-) = \frac{\gamma^{1-\delta}\log\gamma}{\gamma - 1}\,.$$

In the latter case, there exists a unique $0 < \delta^* < 1$ such that $f'_{\delta^*}(0+) = f'_{\delta^*}(1-) = 1$; explicitly $\gamma^{1-\delta^*} = (\gamma - 1)/\log\gamma$, and thus

$$\delta^* = \frac{\log\gamma - \log(\gamma - 1) + \log\log\gamma}{\log\gamma}\,.$$

Consequently, for $0 < \delta < 1$ the graph of f_δ can have three qualitatively different forms.



Note that g_δ always attains its maximum at $x = \delta$. This suggests introducing the auxiliary function ψ according to

$$\psi(\delta) := g_\delta(\delta) = f_\delta(\delta) - \delta = \frac{\gamma - \gamma^{1-\delta}}{\gamma - 1} - \delta\,.$$

It follows from

$$\psi'(\delta) = \frac{\gamma^{1-\delta}\log\gamma}{\gamma - 1} - 1 = \gamma^{\delta^* - \delta} - 1\,,$$

that ψ is concave, with $\psi(0) = \psi(1) = 0$ and $\psi'(\delta^*) = 0$. Hence $\psi(\delta) > 0$ for all $0 < \delta < 1$, and

$$\max\nolimits_{0\leq\delta\leq 1}\psi(\delta) = \psi(\delta^*) = \frac{\gamma - \gamma^{1-\delta^*}}{\gamma - 1} - \delta^* = \frac{\gamma\log\gamma - \gamma + 1}{(\gamma - 1)\log\gamma} - \delta^* = 2C_\gamma\,.$$

Assume first that $0 \le \delta \le \delta^*$. In this case, the function g_δ has a non-positive minimum at $x = 1 + \delta - \delta^* > \delta$, with

$$g_\delta(1+\delta-\delta^*) = 1 - \frac{\gamma^{1-\delta} - \gamma^{1-\delta^*}}{\gamma - 1} - 1 - \delta + \delta^* = \psi(\delta) - \psi(\delta^*)\,,$$

showing that $\max_{0\le x\le 1} g_\delta(x) = \psi(\delta)$ as well as $-\min_{0\le x\le 1} g_\delta(x) = \psi(\delta^*) - \psi(\delta)$. Similarly, if $\delta^* < \delta < 1$ then g_δ has a negative minimum at $x = \delta - \delta^* < \delta$, with

$$g_\delta(\delta - \delta^*) = \frac{\gamma^{1-\delta^*} - \gamma^{1-\delta}}{\gamma - 1} - \delta + \delta^* = \psi(\delta) - \psi(\delta^*)\,.$$

For all $0 \le \delta < 1$, therefore,

$$\max_{0\le x\le 1} g_\delta(x) = \psi(\delta)\,, \quad -\min_{0\le x\le 1} g_\delta(x) = \psi(\delta^*) - \psi(\delta)\,,$$

and consequently

$$\max_{0\le x\le 1} |g_\delta(x)| = \max\bigl\{\psi(\delta), \psi(\delta^*) - \psi(\delta)\bigr\} \ge \tfrac{1}{2}\psi(\delta^*) = C_\gamma\,,$$

which establishes (1) for the case $a = 0$.

To verify (1) for $a > 0$ assume for the time being that $\log_\gamma a = k \in \mathbb{Z}$, and let $l := \lfloor \log_\gamma b \rfloor$ and $\delta := [\![\log_\gamma b]\!]$ as before; for convenience set $m := l - k \in \mathbb{N}_0$. A short computation confirms that

$$f_{m,\delta}(x) := F_{[\![\log_\gamma U_{a,b}]\!]}(x) = \begin{cases} \dfrac{\gamma^x - 1}{\gamma - 1} \cdot \dfrac{\gamma^{m+1} - 1}{\gamma^{m+\delta} - 1} & \text{if } 0 \le x < \delta\,, \\[2ex] 1 - \dfrac{\gamma - \gamma^x}{\gamma - 1} \cdot \dfrac{\gamma^m - 1}{\gamma^{m+\delta} - 1} & \text{if } \delta \le x \le 1\,. \end{cases}$$

Notice that $f_{m,\delta} \to f_\delta$ uniformly on $[0,1]$ as $m \to \infty$. Let again $g_{m,\delta}(x) := f_{m,\delta}(x) - x$ and observe that

$$g_{m,\delta}(x) - g_\delta(x) = f_{m,\delta}(x) - f_\delta(x) = \Delta_{m,\delta}(x)\,,$$

where $\Delta_{m,\delta}$ is given by

$$\Delta_{m,\delta}(x) = \begin{cases} \dfrac{\gamma^x - 1}{\gamma - 1} \cdot \dfrac{\gamma^{1-\delta} - 1}{\gamma^{m+\delta} - 1} & \text{if } 0 \le x < \delta\,, \\[2ex] \dfrac{\gamma^{1-\delta} - \gamma^{x-\delta}}{\gamma - 1} \cdot \dfrac{\gamma^\delta - 1}{\gamma^{m+\delta} - 1} & \text{if } \delta \le x \le 1\,. \end{cases}$$

Obviously, $\Delta_{m,\delta} \ge 0$ with $\Delta_{m,\delta}(0) = \Delta_{m,\delta}(1) = 0$, and $\Delta_{m,0} = 0$ for all $m \ge 1$. Furthermore, for $0 < \delta < 1$ the function $\Delta_{m,\delta}$ is convex and increasing on $[0,\delta]$, and concave and decreasing on $[\delta, 1]$. Since both g_δ and $\Delta_{m,\delta}$ attain their respective maximal value at $x = \delta$,

$$\max_{0\le x\le 1} g_{m,\delta}(x) = g_{m,\delta}(\delta) = g_\delta(\delta) + \Delta_{m,\delta}(\delta) = \psi(\delta) + \Delta_{m,\delta}(\delta)\,.$$

If $0 \le \delta \le \delta^*$ then, with the appropriate $0 \le \xi \le 1$,

$$\begin{aligned} \max_{0\le x\le 1} g_{m,\delta}(x) - \min_{0\le x\le 1} g_{m,\delta}(x) &= g_{m,\delta}(\delta) - g_{m,\delta}(\xi) \\ &\ge g_{m,\delta}(\delta) - g_{m,\delta}(1+\delta-\delta^*) \\ &= \psi(\delta) + \Delta_{m,\delta}(\delta) - g_\delta(1+\delta-\delta^*) - \Delta_{m,\delta}(1+\delta-\delta^*) \\ &= \psi(\delta^*) + \Delta_{m,\delta}(\delta) - \Delta_{m,\delta}(1+\delta-\delta^*) \\ &\ge \psi(\delta^*) \\ &= 2C_\gamma\,. \end{aligned}$$

The same argument applies for $\delta^* < \delta < 1$ with $1 + \delta - \delta^*$ replaced by $\delta - \delta^*$. Thus

$$\max_{0 \le x \le 1} g_{m,\delta}(x) - \min_{0 \le x \le 1} g_{m,\delta}(x) \ge 2C_\gamma \tag{2}$$

holds for all $m \in \mathbb{N}_0$ and $0 \le \delta < 1$, and this in turn implies (1) since

$$\max_{0 \le x \le 1} |g_{m,\delta}(x)| \ge \tfrac{1}{2}\big(\max_{0 \le x \le 1} g_{m,\delta}(x) - \min_{0 \le x \le 1} g_{m,\delta}(x)\big) \ge C_\gamma \, .$$

Overall, therefore, the proof is complete if $a = 0$ or $\log_\gamma a \in \mathbb{Z}$.

Finally, assume that $a > 0$ does not satisfy $\log_\gamma a \in \mathbb{Z}$, that is, the number $\tau := [\![\log_\gamma a]\!]$ lies strictly between 0 and 1. Note that

$$[\![\log_\gamma U_{a,b}]\!] = [\![\log_\gamma U_{a\gamma^{-\tau}, b\gamma^{-\tau}} + \tau]\!] \, ,$$

and clearly $\log_\gamma(a\gamma^{-\tau}) \in \mathbb{Z}$. It is readily verified that, for every non-atomic probability measure μ on $\mathbb{R}$ and every $t \in \mathbb{R}$,

$$F_{[\![\mu+t]\!]}(x) = \begin{cases} F_{[\![\mu]\!]}(x + 1 - [\![t]\!]) - F_{[\![\mu]\!]}(1 - [\![t]\!]) & \text{if } 0 \le x < [\![t]\!] \, , \\ F_{[\![\mu]\!]}(x - [\![t]\!]) + 1 - F_{[\![\mu]\!]}(1 - [\![t]\!]) & \text{if } [\![t]\!] \le x \le 1 \, , \end{cases}$$

and therefore also

$$G_{[\![\mu+t]\!]}(x) = \begin{cases} G_{[\![\mu]\!]}(x + 1 - [\![t]\!]) - G_{[\![\mu]\!]}(1 - [\![t]\!]) & \text{if } 0 \le x < [\![t]\!] \, , \\ G_{[\![\mu]\!]}(x - [\![t]\!]) - G_{[\![\mu]\!]}(1 - [\![t]\!]) & \text{if } [\![t]\!] \le x \le 1 \, , \end{cases}$$

where generally $G_\mu(x) := F_\mu(x) - x$. In particular,

$$\max_{0 \le x \le 1} G_{[\![\mu+t]\!]}(x) - \min_{0 \le x \le 1} G_{[\![\mu+t]\!]}(x) = \max_{0 \le x \le 1} G_{[\![\mu]\!]}(x) - \min_{0 \le x \le 1} G_{[\![\mu]\!]}(x) \, , \tag{3}$$

which merely expresses the intuitively obvious fact that the *span* (i.e. the difference between maximal and minimal value) of $G_{[\![\mu]\!]}$ is not affected by the rotation caused by adding (modulo one) any number t. With the notation introduced earlier, $G_{[\![\log_\gamma U_{a\gamma^{-\tau}, b\gamma^{-\tau}}]\!]} = g_{m,\delta}$, where $m = \lfloor \log_\gamma b/a \rfloor$ and $\delta = [\![\log_\gamma b/a]\!]$. Combining (2) and (3) for $\mu = \log_\gamma U_{a\gamma^{-\tau}, b\gamma^{-\tau}}$ and $t = \tau$ therefore yields

$$\max_{0 \le x \le 1} G_{[\![\log_\gamma U_{a,b}]\!]}(x) - \min_{0 \le x \le 1} G_{[\![\log_\gamma U_{a,b}]\!]}(x) \ge 2C_\gamma \, .$$

This completes the proof. □

Remark 2. (i) As the above argument shows, the constant C_γ in (1) is best possible: For every $C > C_\gamma$ there exist a, b with $0 < a < b$ such that $d_\infty([\![\log_\gamma U_{a,b}]\!], U_{0,1}) < C$.

(ii) It follows from the first part of the proof of Theorem 1 that, for all $\gamma > 1$ and $b > 0$,

$$d_\infty\big([\![\log_\gamma U_{0,b}]\!], U_{0,1}\big) = \Psi(\log_\gamma b) \, ,$$

with the continuous, 1-periodic function $\Psi : x \mapsto \max\big\{\psi([\![x]\!]), 2C_\gamma - \psi([\![x]\!])\big\} = C_\gamma + |\psi([\![x]\!]) - C_\gamma|$.

(iii) Note that $\gamma \mapsto C_\gamma$ is monotonically increasing, with $\lim_{\gamma \to 1+} C_\gamma = 0$ and $\lim_{\gamma \to \infty} C_\gamma = \frac{1}{2}$. For $\gamma = 10$, the most important special case in view of Benford's Law, one finds $C_{10} \approx 0.13442$.

(iv) Satisfactory though it may be, Theorem 1 has a small shortcoming: For every probability measure μ on $\mathbb{R}$, the measure $[\![\mu]\!]$ naturally lives on $\mathbb{T} = \mathbb{R}/\mathbb{Z}$ rather than on $[0, 1]$, but d_∞ is unsuitable for measures on $\mathbb{T}$. Specifically, $\mathbb{T}$ is a compact metric space when endowed with the metric $d(x + \mathbb{Z}, y + \mathbb{Z}) := \min_{k \in \mathbb{Z}} |x - y + k|$, and consequently $\mathcal{P}(\mathbb{T})$, the space of all probability measures on $\mathbb{T}$ with the topology of weak convergence, is compact and metrizable [3]. A natural metric inducing this topology is the Kantorovich–Wasserstein distance d_K defined as

$$d_K(\mu, \nu) := \sup\left\{ \left| \int_{\mathbb{T}} f \, \mathrm{d}\mu - \int_{\mathbb{T}} f \, \mathrm{d}\nu \right| \, : \, f \in C_{\mathbb{R}}(\mathbb{T}), \mathrm{Lip} f \le 1 \right\} \, .$$

Unlike d_∞, the metric d_K on $\mathcal{P}(\mathbb{T})$ is invariant under isometries of $\mathbb{T}$: If $T : \mathbb{T} \to \mathbb{T}$ is any isometry, then $d_K(T\mu, T\nu) = d_K(\mu, \nu)$ for all $\mu, \nu \in \mathcal{P}(\mathbb{T})$. Explicit practicable formulae for d_K have been derived in [1]. A truly satisfactory variant of Theorem 1, therefore, would consider $[\![\log_\gamma U_{a,b}]\!]$ an element of $\mathcal{P}(\mathbb{T})$ and provide a lower bound for its distance from $\lambda_\mathbb{T}$, the uniform distribution on $\mathbb{T}$. Such a result can indeed be achieved using parts of the proof of Theorem 1 even though the necessary calculations are significantly more involved. The final result, however, is even slightly simpler than (1): For all $\gamma > 1$ and $0 \le a < b$,

$$d_K\big([\![\log_\gamma U_{a,b}]\!], \lambda_\mathbb{T}\big) \ge \log_\gamma \tfrac{1+\sqrt{\gamma}}{2} - \tfrac{1}{4} =: \tfrac{1}{4}\Phi(\tfrac{1}{4}\log\gamma) > 0\,, \tag{4}$$

where Φ is the real-analytic odd function $\Phi(x) = x^{-1}\log\cosh x$; as in the case of (1), the inequality (4) is best possible in the sense of (i).

References

[1] C.A. Cabrelli and U.M. Molter, The Kantorovich metric for probability measures on the circle, *J. Comput. Appl. Math.* **57** (1995), 345–361.

[2] P. Diaconis, The distribution of leading digits and uniform distribution mod 1, *Ann. Probab.* **5** (1979), 72–81.

[3] R.M. Dudley, *Real analysis and probability*, Revised reprint of the 1989 original, Cambridge University Press (2002).

[4] W. Feller, *An introduction to probability theory and its applications*, Vol. II. 2nd ed., John Wiley and Sons, New York (1971).

[5] A.L. Gibbs and F.E. Su, On choosing and bounding probability metrics, *Int. Stat. Rev.* **70** (2002), 419–435.