J Theor Probab
DOI 10.1007/s10959-007-0112-z

# Scale-Distortion Inequalities for Mantissas of Finite Data Sets

**Arno Berger · Theodore P. Hill · Kent E. Morrison**

Received: 10 June 2007 / Revised: 21 August 2007

**Abstract** In scientific computations using floating point arithmetic, rescaling a data set multiplicatively (e.g., corresponding to a conversion from dollars to euros) changes the distribution of the mantissas, or fraction parts, of the data. A scale-distortion factor for probability distributions is defined, based on the Kantorovich distance between distributions. Sharp lower bounds are found for the scale-distortion of $n$-point data sets, and the unique data set of size $n$ with the least scale-distortion is identified for each positive integer $n$. A sequence of real numbers is shown to follow Benford's Law (base $b$) if and only if the scale-distortion (base $b$) of the first $n$ data points tends zero as $n$ goes to infinity. These results complement the known fact that Benford's Law is the unique scale-invariant probability distribution on mantissas.

**Keywords** Benford's Law · Scale-invariance · Scale-distortion · Mantissa distribution · Kantorovich metric

The first author was partly supported by a HUMBOLDT research fellowship. The second author was supported in part by the National Security Agency and as a Research Scholar in Residence at California Polytechnic State University.

A. Berger
Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand
e-mail: arno.berger@canterbury.ac.nz

T.P. Hill
School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332, USA
e-mail: hill@math.gatech.edu

K.E. Morrison (✉)
Department of Mathematics, California Polytechnic State University, San Luis Obispo,
CA 93407, USA
e-mail: kmorriso@calpoly.edu

## 1 Introduction

In analyzing real-valued numerical data, it is important not only to study the distribution of the raw data itself, but also to study the distribution of the mantissas of the data. For example, as Knuth states in *The Art of Computer Programming* [12, pp. 238], "In order to analyze the average behavior of floating-point arithmetic algorithms (and in particular to determine their average running time), we need some statistical information that allows us to determine how often various cases arise." The decision to terminate an algorithm is often based on the observed values of the mantissas of the output—for example, to stop if $n$ values in a row are identical, or if the difference between successive values is less than a given amount. Thus the running time of the algorithm depends on the empirical distribution of the mantissas. As another example, the analysis of mantissas via goodness-of-fit tests to *Benford's Law*, the well-known logarithmic probability distribution on mantissas, is now widely used for fraud detection, for tests of homogeneity of data, and for diagnostic tests of mathematical models [11, 14].

In general, however, the distribution of both the raw data and the mantissas of the data depends on the units used—converting from dollars to euros, or from meters to feet, will almost always alter the distributions. It is an easy fact that no finite set of mantissas is exactly invariant under arbitrary changes of scale, and it is one of the goals of this article to establish sharp inequalities and bounds on how close to scale-invariant a data set of size $n$ can be, and to identify the data sets altered the least by changes of scale.

Using the classical Kantorovich metric for the distance between probability distributions on a bounded set (the mantissas), a natural scale-distortion factor for distributions of mantissas is defined. For each positive integer $n$, a sharp lower bound is found for the scale-distortion of every $n$-point data set, and the unique most scale-invariant (i.e, least scale-distorted) set of size $n$ is identified (Theorem 3.22). These extremal data sets are then compared with the $n$-point data sets (Corollary 2.10) that are closest to the unique scale-invariant distribution, Benford's logarithmic distribution. These inequalities are used to show that the mantissas of a sequence of real numbers are Benford-distributed if and only if the scale-distortion of the first $n$ points goes to zero as $n$ goes to infinity (Theorem 3.19), from which it follows that the scale-distortion of a sequence of i.i.d. random variables with mantissa distribution $P$ approaches zero almost surely as $n$ goes to infinity, if $P$ is Benford's Law, and if not, then the lim sup of the successive scale-distortions is almost surely strictly positive (Theorem 3.21).

## 2 Notation and Basic Tools

Throughout this article, $b$ denotes a natural number greater than 1, referred to as the *base*. For every $t \in \mathbb{R}^+$, $\langle t\rangle_b$ is the (base $b$) mantissa of $t$, i.e., $\langle t\rangle_b$ is the unique number $u \in [1, b)$ with $t = ub^k$ for some $k \in \mathbb{Z}$.

*Example 2.1* $\langle 71\rangle_{10} = \langle 7.1\rangle_{10} = \langle 0.71\rangle_{10} = 7.1$.

Given a data set $X = \{x_1, \dots, x_n\}$ of points in $\mathbb{R}^+$, i.e., $X$ is an unordered $n$-tuple of positive real numbers, possibly with repetitions, define the probability measures

$$P_X = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i} \quad \text{and} \quad \langle P_X \rangle_b = \frac{1}{n}\sum_{i=1}^{n} \delta_{\langle x_i \rangle_b},$$

where $\delta_t$ denotes the probability measure concentrated at $t \in \mathbb{R}$. Note that $\langle P_X \rangle_b([1, b)) = 1$.

The next definition recalls one of the best-known probability distributions on mantissas, namely Benford's Law [2, 11, 13], which will play a special role in the scale-distortion inequalities below, essentially since it is known to be the unique scale-invariant probability distribution on mantissas [10]. (It is also known to be the unique atomless base-invariant and the unique sum-invariant distribution [1, 10].)

**Definition 2.2** A sequence $(x_n)$ of positive real numbers is *b-Benford* (or *Benford base b*) if

$$\lim_{n\to\infty} \frac{\#\{i \le n : \langle x_i \rangle_b \le t\}}{n} = \log_b t \quad \text{for all } t \in [1, b).$$

Inherent in Definition 2.2 is *Benford's Law*, the Borel probability measure $\mathbb{B}_b$ on $\mathbb{R}^+$ with

$$\mathbb{B}_b([1, t]) = \log_b t \quad \text{for all } t \in [1, b).$$

Obviously, $\mathbb{B}_b([1, b)) = 1$. (Here and throughout, the symbol $\log_b$ denotes the logarithm base $b$; if used without a subscript, log means the natural logarithm.)

Recall that a sequence $(P_n)$ of probability measures on $\mathbb{R}$, with associated distribution functions (d.f.'s) $F_{P_n}$, converges weakly to $P$, with d.f. $F_P$, if and only if $(F_{P_n})$ converges pointwise to $F_P$ at every point of continuity of $F_P$.

**Proposition 2.3** *The sequence $(x_n)$ of positive real numbers is b-Benford if and only if $\langle P_{X_n} \rangle_b \to \mathbb{B}_b$ weakly as $n \to \infty$, where $X_n = \{x_1, \dots, x_n\}$ for each $n \in \mathbb{N}$.*

*Proof* Let $F_n$ be the d.f. of $\langle P_{X_n} \rangle_b$. Then

$$F_n(t) = \frac{\#\{i \le n : \langle x_i \rangle_b \le t\}}{n},$$

and $\langle P_{X_n} \rangle_b \to \mathbb{B}_b$ weakly if and only if $F_n(t) \to \log_b(t)$ for all $t \in [1, b)$, that is, if and only if $(x_n)$ is $b$-Benford. □

Let $\mathcal{P}(\mathbb{R})$ denote the family of all Borel probability measures on $\mathbb{R}$. It is well-known that, with the topology of weak convergence, $\mathcal{P}(\mathbb{R})$ can be given the structure of a complete, separable metric space in different ways, that is, by means of different metrics. For the practical purpose of quantifying scale-distortion an easily computed metric is required. Since mantissas are bounded, it is enough to consider probability measures with finite expectation only, i.e., to restrict to the subset

$$\mathcal{P}_1(\mathbb{R}) := \left\{ P \in \mathcal{P}(\mathbb{R}) : \int_{\mathbb{R}} |t| \, dP(t) < \infty \right\}$$

of $\mathcal{P}(\mathbb{R})$. For every $P \in \mathcal{P}(\mathbb{R})$ denote by $\operatorname{supp} P$ its support, i.e., $\operatorname{supp} P$ is the smallest closed set with $P$-measure 1. Clearly $P \in \mathcal{P}_1(\mathbb{R})$ whenever $\operatorname{supp} P$ is compact. If $F_P$ is the d.f. of $P \in \mathcal{P}(\mathbb{R})$ then, by Fubini's theorem,

$$P \in \mathcal{P}_1(\mathbb{R}) \quad \text{if and only if} \quad \int_{-\infty}^{0} F_P(t) \, dt + \int_{0}^{\infty} (1 - F_P(t)) \, dt < \infty.$$

Let $P_1, P_2 \in \mathcal{P}_1(\mathbb{R})$ with d.f.'s $F_{P_1}, F_{P_2}$. Recall that the *Kantorovich* (or *Wasserstein*) *metric* $d_K$ is defined by

$$d_K(P_1, P_2) = \int_{-\infty}^{\infty} |F_{P_1}(t) - F_{P_2}(t)| \, dt.$$

Given any d.f. $F$, let $F^{-1}: (0, 1) \to \mathbb{R}$ denote its generalized upper inverse (or upper quantile) function, that is, $F^{-1}(t) = \sup\{u : F(u) \le t\}$. Note that, again by Fubini's theorem,

$$\int_{-\infty}^{\infty} |F_{P_1}(t) - F_{P_2}(t)| \, dt = \int_{0}^{1} |F_{P_1}^{-1}(t) - F_{P_2}^{-1}(t)| \, dt. \tag{2.1}$$

There are at least three reasons for choosing the Kantorovich distance as a means to quantify scale-distortion. First, it is easy to compute, unlike the Lévy and Prokhorov metrics. Second, it is a *bona fide* metric and metrizes weak convergence on spaces of bounded diameter (see Lemma 2.6 below). Third, it has a clear intuitive probabilistic interpretation: By the celebrated Kantorovich-Rubinstein theorem [8, Theorem 11.8.2], it is the minimal expected distance between two jointly distributed random variables $\xi_1, \xi_2$ with marginals $P_1$ and $P_2$, respectively, that is,

$$d_K(P_1, P_2) = \inf\{\mathbb{E}|\xi_1 - \xi_2| : \mathcal{L}(\xi_1) = P_1, \, \mathcal{L}(\xi_2) = P_2, \, \xi_1, \xi_2 \text{ jointly distributed}\}, \tag{2.2}$$

where $\mathcal{L}(\xi)$ denotes the law, or probability distribution, of the random variable $\xi$.

*Example 2.4* Let $P$ be uniform on $[1, b)$. Then

$$d_K(P, \mathbb{B}_b) = \int_{1}^{b} \left( \log_b t - \frac{t-1}{b-1} \right) dt = \frac{b+1}{2} - \frac{b-1}{\log b} > 0.$$

*Example 2.5* Let $b = 10$, $X = \{1, 2\}$, $Y = \{2, 3\}$, $Z = \{1, 2, 3\}$. Then

$$d_K(\langle P_X \rangle_b, \langle P_Y \rangle_b) = 1, \qquad d_K(\langle P_X \rangle_b, \langle P_Z \rangle_b) = 1/2,$$
$$d_K(\langle P_Y \rangle_b, \langle P_Z \rangle_b) = 1/2.$$

That the Kantorovich metric is truly a metric and that it metrizes weak convergence of probability measures on spaces of bounded diameter is known [8, 9]; a proof of these facts for the special case of probability measures on mantissas is included for completeness. Denote by $\mathcal{P}[1,b)$ the set of Borel probability measures on $[1,b)$, that is,

$$\mathcal{P}[1,b) = \{P \in \mathcal{P}(\mathbb{R}) : P([1,b)) = 1\},$$

and recall that a metric $d(\cdot,\cdot)$ on a space of probability measures $\mathcal{S}$ *metrizes weak convergence on* $\mathcal{S}$ if, for all $P \in \mathcal{S}$ and all sequences $(P_n)$ in $\mathcal{S}$, $d(P, P_n) \to 0$ if and only if $P_n \to P$ weakly.

**Lemma 2.6** *For all* $b \in \mathbb{N} \setminus \{1\}$:

(i) $d_K$ *is a metric on* $\mathcal{P}[1,b)$;
(ii) $d_K$ *metrizes weak convergence on* $\mathcal{P}[1,b)$.

*Proof* (i) Obviously, $\mathcal{P}[1,b) \subset \mathcal{P}_1(\mathbb{R})$, hence $d_K(P_1, P_2) < \infty$ for any two $P_1, P_2 \in \mathcal{P}[1,b)$. The right-continuity of d.f.'s implies that two d.f.'s that agree almost everywhere are identical. Thus, the standard one-to-one correspondence between Borel probability measures $P \in \mathcal{P}[1,b)$ and d.f.'s $F$ on $[1,b)$ (i.e., $F$ is non-decreasing and right-continuous with $F(1) \ge 0$ and $\lim_{t\uparrow b} F(t) = 1$, see e.g. [6, Theorem 2.2.4]) implies that $\mathcal{P}[1,b)$ may be identified via $P \mapsto F_P$ with a subset of $L_1[1,b)$, the space of $L_1$-functions on $[1,b)$. Hence $d_K$ is simply the standard $L_1$-metric on $L_1[1,b)$, restricted to the set of d.f.'s.

(ii) Let $d_P$ denote the Prokhorov metric on $\mathcal{P}[1,b)$ (cf. [8]), that is,

$$d_P(P_1, P_2) = \inf\{\varepsilon > 0 : P_1(B) \le P_2(B^\varepsilon) + \varepsilon \text{ for all Borel subsets } B \text{ of } [1,b)\},$$

where

$$B^\varepsilon = \{t \in [1,b) : \inf_{u\in B} |u - t| < \varepsilon\}.$$

By [9, Theorem 2],

$$(d_P)^2 \le d_K \le b\, d_P,$$

and since $d_P$ metrizes weak convergence on any separable metric space (e.g., [8, p. 81]), this implies that $d_K$ metrizes weak convergence on $\mathcal{P}[1,b)$. □

Recall that $\langle P_X\rangle_b \ne \mathbb{B}_b$ for every finite data set $X$. To quantify how small $d_K(\langle P_X\rangle_b, \mathbb{B}_b)$ can be for a data set $X$ of size $n$, it is helpful to address the following more general question: Given $P \in \mathcal{P}_1(\mathbb{R})$, what is the smallest possible value of $d_K(P, \frac{1}{n}\sum_{i=1}^n \delta_{x_i})$, where $x_1, \ldots, x_n \in \mathbb{R}$? This question will be answered completely in Theorem 2.8 below; for $n = 1$ the latter reduces to the well-known fact [4, p. 54] that, for any integrable real-valued random variable $\xi$,

$$\mathbb{E}(|\xi - x_1|) \text{ is minimal} \quad \Longleftrightarrow \quad x_1 \text{ is a median of } \xi. \tag{2.3}$$

Generally, given $P \in \mathcal{P}(\mathbb{R})$ with corresponding d.f. $F_P$ and $t \in (0,1)$, the *$t$-quantile set* $I_t^P$ of $P$ is defined as

$$I_t^P = \big[\inf\{u : F_P(u) \ge t\}, \sup\{u : F_P(u) \le t\}\big].$$

The following lemma records several well-known useful facts about quantile sets; proofs are included for the sake of completeness.

**Lemma 2.7** *Let $P \in \mathcal{P}(\mathbb{R})$ with d.f. $F_P$. Then, for every $t \in (0,1)$:*

(i) $I_t^P$ *is a non-empty, compact (possibly one-point) interval* $[\alpha, \beta]$;
(ii) $\{\alpha, \beta\} \subset \operatorname{supp} P$ *and* $(\alpha, \beta) \subset \mathbb{R}\backslash \operatorname{supp} P$;
(iii) $F_P((\alpha, \beta)) \subset \{t\}$.

*Furthermore, if $t_1 < t_2$ then $u \le v$ for every $u \in I_{t_1}^P$ and every $v \in I_{t_2}^P$, and $I_{t_1}^P \cap I_{t_2}^P$ contains at most one point.*

*Proof* Fix $t \in (0,1)$ and let $\alpha = \inf\{u : F_P(u) \ge t\}$, $\beta = \sup\{u : F_P(u) \le t\}$.

(i) Since $F_P$ is non-decreasing with $\lim_{u\to-\infty} F_P(u) = 0$ and $\lim_{u\to\infty} F_P(u) = 1$, both $\alpha$ and $\beta$ are finite. Moreover, $F_P(u) < t$ whenever $u < \alpha$ and thus $\beta \ge u$. Consequently, $\beta \ge \alpha$, and $I_t^P = [\alpha, \beta]$ is a non-empty, compact interval.

(ii) Suppose $F_P(\alpha - \varepsilon) = F_P(\alpha)$ for some $\varepsilon > 0$. Then $F_P(\alpha - \varepsilon) = F_P(\alpha) \ge t$, an obvious contradiction to the definition of $\alpha$. Therefore $\alpha \in \operatorname{supp} P$. Similarly, if $F_P(\beta) = F_P(\beta + \varepsilon)$ for some $\varepsilon > 0$ then $P(\{\beta\}) > 0$ because otherwise $F_P(\beta + \varepsilon) \le t$, which clearly contradicts the definition of $\beta$. Hence, $\{\alpha, \beta\} \subset \operatorname{supp} P$. For any $u$ with $\alpha < u < \beta$ clearly $F_P(u) = t$, implying that $u \in \mathbb{R}\backslash \operatorname{supp} P$.

(iii) This is obvious from part (ii).

To conclude the proof of the lemma, let $t_1 < t_2$ and pick any $u \in I_{t_1}^P$, $v \in I_{t_2}^P$. If $u > v$ then $F_P(\frac{1}{2}(u+v)) \ge F_P(v) \ge t_2$ and so $\lim_{w\uparrow u} F_P(w) \ge t_2$, which is impossible. Thus $u \le v$. If $u \in I_{t_1}^P \cap I_{t_2}^P$ and $v > u$ then $\lim_{w\uparrow v} F_P(w) \ge F_P(u) \ge t_2$, and so $v \notin I_{t_1}^P$. Analogously, if $v < u$ then $F_P(v) \le \lim_{w\uparrow u} F_P(w) \le t_1$, so $v \notin I_{t_2}^P$. Hence, $I_{t_1}^P \cap I_{t_2}^P = \{u\}$ and $P(\{u\}) \ge t_2 - t_1 > 0$. □

Given a random variable $\xi$ with $\mathcal{L}(\xi) = P$ and a one-point data set $X = \{x_1\}$, (2.2) implies that an equivalent form of (2.3) is

$$d_K(P, P_X) \text{ is minimal} \quad \Longleftrightarrow \quad x_1 \in I_{1/2}^P. \tag{2.4}$$

The following theorem, the main theorem of this section, generalizes (2.4) to arbitrary finite data sets $X$. This result will be used in the next section to show that the $n$-point data set having the least scale-distortion is *not* the same as—although a scaled version of—the $n$-point data set closest (w.r.t. the Kantorovich metric) to the unique scale-invariant distribution $\mathbb{B}_b$.

**Theorem 2.8** *Let $P \in \mathcal{P}_1(\mathbb{R})$ and $n \in \mathbb{N}$. For the data set $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}$ with $x_1 \le \cdots \le x_n$ the distance $d_K(P, P_X)$ is minimal if and only if $x_i \in I_{(2i-1)/(2n)}^P$ for all $i = 1, \ldots, n$.*

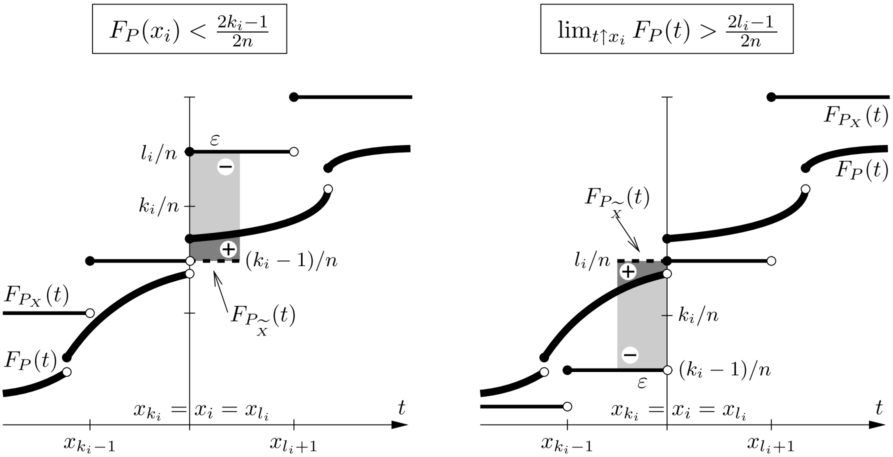

**Fig. 1** If $F_P(x_i) < \frac{2i-1}{2n}$ or if $\lim_{t\uparrow x_i} F_P(t) > \frac{2i-1}{2n}$ then $d_K(P, P_X)$ is not minimal. The *shaded areas* illustrate the net decrease in $d_K(P, P_X)$ if some $x_j$ are moved slightly to the *right* or *left*, respectively

*Proof* Assume that $X$ is a data set of size $n$ such that $d_K(P, P_X)$ is minimal. First, suppose that there is some $i \in \{1, 2, \dots, n\}$ such that $F_P(x_i) < \frac{2i-1}{2n}$ and let

$$k_i = \min\left\{1 \le k \le i : x_k = x_i,\ F_P(x_i) < \frac{2k-1}{2n}\right\}$$

and also

$$l_i = \max\{i \le l \le n : x_l = x_i\},$$

so that in particular $1 \le k_i \le i \le l_i \le n$. Since $F_P$ is right-continuous, there exists $\varepsilon_1 > 0$ such that

$$\frac{2k_i-3}{2n} \le F_P(t) < \frac{2k_i-1}{2n} \quad \text{for all } t \in [x_i, x_i + \varepsilon_1],$$

and hence

$$\left|F_P(t) - \frac{k_i-1}{n}\right| \le \frac{1}{2n} \quad \text{for all } t \in [x_i, x_i + \varepsilon_1]. \tag{2.5}$$

If $l_i = n$ let $\varepsilon = \varepsilon_1$, otherwise let $\varepsilon = \min(\varepsilon_1, \frac{1}{2}(x_{l_i+1} - x_i))$, and consider the $n$-point data set

$$\widetilde{X} = \{x_1, \dots, x_{k_i-1}, x_{k_i} + \varepsilon, \dots, x_{l_i} + \varepsilon, x_{l_i+1}, \dots, x_n\},$$

i.e., $\widetilde{X}$ is created from $X$ by moving $x_{k_i}, \dots, x_{l_i}$ slightly to the *right*, see also Fig. 1. Clearly, $F_{P_{\widetilde{X}}}(t) = F_{P_X}(t)$ whenever $t \notin [x_i, x_i + \varepsilon]$. Then

$$\begin{aligned} &d_K(P, P_X) - d_K(P, P_{\widetilde{X}}) \\ &\quad = \int_{-\infty}^{\infty} |F_P(t) - F_{P_X}(t)|\,dt - \int_{-\infty}^{\infty} |F_P(t) - F_{P_{\widetilde{X}}}(t)|\,dt \end{aligned}$$

$$
\begin{aligned}
&= \int_{x_i}^{x_i+\varepsilon} \big( |F_P(t) - F_{P_X}(t)| - |F_P(t) - F_{P_{\widetilde{X}}}(t)| \big)\, dt \\
&= \int_{x_i}^{x_i+\varepsilon} \left( \frac{l_i}{n} - F_P(t) - \left| F_P(t) - \frac{k_i - 1}{n} \right| \right) dt \\
&\geq \int_{x_i}^{x_i+\varepsilon} \left( \frac{2l_i - 1}{2n} - F_P(t) \right) dt \geq \int_{x_i}^{x_i+\varepsilon} \left( \frac{2k_i - 1}{2n} - F_P(t) \right) dt > 0,
\end{aligned}
$$

where the last two weak inequalities follow from (2.5) together with $l_i \geq k_i$. This implies that $d_K(P, P_X) > d_K(P, P_{\widetilde{X}})$, contradicting the minimality of $d_K(P, P_X)$. Hence $F_P(x_i) \geq \frac{2i-1}{2n}$.

The argument for the case that $\lim_{t\uparrow x_i} F_P(t) > \frac{2i-1}{2n}$ is analogous but slightly different because of the right-continuity of distribution functions. In this case let

$$k_i = \min\{1 \leq k \leq i : x_k = x_i\},$$

and

$$l_i = \max\left\{ i \leq l \leq n : \lim_{t\uparrow x_i} F_P(t) > \frac{2l - 1}{2n} \right\},$$

so that again $1 \leq k_i \leq i \leq l_i \leq n$. There now exists $\varepsilon_1 > 0$ such that

$$\frac{2l_i - 1}{2n} < F_P(t) \leq \frac{2l_i + 1}{2n} \quad \text{for all } t \in [x_i - \varepsilon_1, x_i),$$

and thus

$$\left| F_P(t) - \frac{l_i}{n} \right| \leq \frac{1}{2n} \quad \text{for all } t \in [x_i - \varepsilon_1, x_i).$$

If $k_i = 1$ let $\varepsilon = \varepsilon_1$, otherwise let $\varepsilon = \min(\varepsilon_1, \frac{1}{2}(x_i - x_{k_i - 1}))$, and consider the $n$-point data set

$$\widetilde{X} = \{x_1, \ldots, x_{k_i - 1}, x_{k_i} - \varepsilon, \ldots, x_{l_i} - \varepsilon, x_{l_i + 1}, \ldots, x_n\},$$

i.e., $\widetilde{X}$ is created from $X$ by moving $x_{k_i}, \ldots, x_{l_i}$ slightly to the *left* (cf. Fig. 1). Clearly, $F_{P_{\widetilde{X}}}$ and $F_{P_X}$ coincide outside $[x_i - \varepsilon, x_i]$, and

$$
\begin{aligned}
&d_K(P, P_X) - d_K(P, P_{\widetilde{X}}) \\
&\quad = \int_{-\infty}^{\infty} |F_P(t) - F_{P_X}(t)|\, dt - \int_{-\infty}^{\infty} |F_P(t) - F_{P_{\widetilde{X}}}(t)|\, dt \\
&\quad = \int_{x_i - \varepsilon}^{x_i} \left( \left| F_P(t) - \frac{k_i - 1}{n} \right| - \left| F_P(t) - \frac{l_i}{n} \right| \right) dt \\
&\quad = \int_{x_i - \varepsilon}^{x_i} \left( F_P(t) - \frac{k_i - 1}{n} - \left| F_P(t) - \frac{l_i}{n} \right| \right) dt \\
&\quad \geq \int_{x_i - \varepsilon}^{x_i} \left( F_P(t) - \frac{2k_i - 1}{2n} \right) dt \geq \int_{x_i - \varepsilon}^{x_i} \left( F_P(t) - \frac{2l_i - 1}{2n} \right) dt > 0,
\end{aligned}
$$

so that $d_K(P, P_X) > d_K(P, P_{\widetilde{X}})$, again contradicting the minimality of $d_K(P, P_X)$. Hence $\lim_{t\uparrow x_i} F_P(t) \le \frac{2i-1}{2n}$. Overall therefore

$$\lim_{t\uparrow x_i} F_P(t) \le \frac{2i-1}{2n} \le F_P(x_i) \quad \text{for all } i = 1, \ldots, n,$$

or, equivalently,

$$x_i \in I^P_{(2i-1)/(2n)} \quad \text{for all } i = 1, \ldots, n, \tag{2.6}$$

whenever $d_K(P, P_X)$ is minimal for $X = \{x_1, \ldots, x_n\}$.

For the converse, assume that (2.6) holds, let $\Delta_n = \{x \in \mathbb{R}^n : x_1 \le \cdots \le x_n\}$, and consider the non-negative function

$$\varphi : \begin{cases} \Delta_n \to \mathbb{R}, \\ x \mapsto d_K(P, P_X), \end{cases} \qquad \text{where } X = \{x_1, \ldots, x_n\}.$$

Endow $\Delta_n$ with a metric induced by any norm on $\mathbb{R}^n$ (e.g. the $\ell_1$-norm, see Proposition 2.12 below). It is easy to check that $\varphi$ is Lipschitz continuous, and $\varphi(x) \to \infty$ as $x_1 \to -\infty$ or $x_n \to \infty$. Hence $\varphi$ attains a minimum, say at $y = (y_1, \ldots, y_n) \in \Delta_n$. Fix $i \in \{1, 2, \ldots, n\}$ and note that $y_i \in I^P_{(2i-1)/(2n)}$. Let $x_1 \le \cdots \le x_n$ satisfy (2.6). If $x_i \ne y_i$ then $I^P_{(2i-1)/(2n)}$ is not a singleton, and so $F_P(t) = \frac{2i-1}{2n}$ for every $t$ in the interior of $I^P_{(2i-1)/(2n)}$. Let $I^P_{(2i-1)/(2n)} = [\alpha, \beta]$ and consider the data set $\widetilde{X} = \{x_1, \ldots, x_{i-1}, y_i, x_{i+1}, \ldots, x_n\}$. Clearly, $F_{P_{\widetilde{X}}}$ and $F_{P_X}$ coincide outside $I^P_{(2i-1)/(2n)}$. From

$$\begin{aligned} & d_K(P, P_X) - d_K(P, P_{\widetilde{X}}) \\ &\quad = \int_{I^P_{(2i-1)/(2n)}} |F_P(t) - F_{P_X}(t)|\, dt - \int_{I^P_{(2i-1)/(2n)}} |F_P(t) - F_{P_{\widetilde{X}}}(t)|\, dt \\ &\quad = \int_{I^P_{(2i-1)/(2n)}} \left| \frac{2i-1}{2n} - F_{P_X}(t) \right| dt - \int_{I^P_{(2i-1)/(2n)}} \left| \frac{2i-1}{2n} - F_{P_{\widetilde{X}}}(t) \right| dt \\ &\quad = \int_\alpha^{x_i} \left| \frac{2i-1}{2n} - \frac{i-1}{n} \right| dt + \int_{x_i}^{\beta} \left| \frac{2i-1}{2n} - \frac{i}{n} \right| dt \\ &\qquad - \int_\alpha^{y_i} \left| \frac{2i-1}{2n} - \frac{i-1}{n} \right| dt - \int_{y_i}^{\beta} \left| \frac{2i-1}{2n} - \frac{i}{n} \right| dt = 0, \end{aligned}$$

it follows that $\varphi(x_1, \ldots, x_{i-1}, y_i, x_{i+1}, \ldots, x_n) = \varphi(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n)$. Since $i$ was arbitrary, it follows that $\varphi(x) = \varphi(y)$. Thus $\varphi(x) = d_K(P, P_X)$ is minimal. □

**Corollary 2.9** *Let $P \in \mathcal{P}_1(\mathbb{R})$, $n \in \mathbb{N}$, and $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}$ with $x_1 \le x_2 \le \cdots \le x_n$. If $P$ has no atoms (i.e., $F_P$ is continuous) then $d_K(P, P_X)$ is minimal if and only if $F_P(x_i) = \frac{2i-1}{2n}$ for all $i = 1, \ldots, n$. If* supp $P = \mathbb{R}$ *then the data set $X$ minimizing $d_K(P, P_X)$ is unique.*

*Proof* If $F_P$ is continuous at $x_i$ then $x_i \in I_t^P$ if and only if $F_P(x_i) = t$. By Lemma 2.7(i) and (ii), every quantile set is a singleton if $\operatorname{supp} P = \mathbb{R}$. In particular, $X$ is unique in this case. □

The next corollary identifies the unique $n$-point mantissa data set in $[1, b)$ that is closest in the Kantorovich metric to the unique scale-invariant mantissa distribution $\mathbb{B}_b$, and it identifies the minimal distance. As will be seen in the next section, this unique set is *not* the same as the $n$-point data set having the least scale-distortion.

**Corollary 2.10** *Let $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^+$ be a finite data set. Then*

$$d_K(\langle P_X \rangle_b, \mathbb{B}_b) \geq \frac{b-1}{\log b} \cdot \frac{b^{1/(2n)} - 1}{b^{1/(2n)} + 1} = \frac{b-1}{\log b} \tanh\left(\frac{\log b}{4n}\right). \tag{2.7}$$

*Equality holds in (2.7) if and only if $\{\langle x_1 \rangle_b, \ldots, \langle x_n \rangle_b\} = \{b^{(2i-1)/(2n)} : i = 1, \ldots, n\}$.*

*Proof* Since $F_{\mathbb{B}_b}$ is continuous and strictly increasing, $I_t^{\mathbb{B}_b}$ is the singleton $\{b^t\}$ for each $t \in (0, 1)$. Thus, equality is attained if and only if $\{\langle x_1 \rangle_b, \ldots, \langle x_n \rangle_b\} = \{b^{(2i-1)/(2n)} : i = 1, \ldots, n\}$. Consequently, a straightforward computation yields

$$\begin{aligned}
&d_K\left(\frac{1}{n}\sum_{i=1}^{n} \delta_{b^{(2i-1)/(2n)}}, \mathbb{B}_b\right) \\
&\quad = \int_0^{b^{1/(2n)}} \log_b t \, dt + \sum_{i=1}^{n-1} \int_{b^{(2i-1)/(2n)}}^{b^{(2i+1)/(2n)}} \left|\log_b t - \frac{i}{n}\right| dt \\
&\qquad + \int_{b^{(2n-1)/(2n)}}^{b} (1 - \log_b t) \, dt \\
&\quad = \int_0^{b^{1/(2n)}} \log_b t \, dt + \int_{b^{-1/(2n)}}^{b^{1/(2n)}} |\log_b t| \, dt \sum_{i=1}^{n-1} b^{i/n} \\
&\qquad + \int_{b^{(2n-1)/(2n)}}^{b} (1 - \log_b t) \, dt \\
&\quad = \frac{b-1}{\log b} \cdot \frac{b^{1/(2n)} - 1}{b^{1/(2n)} + 1} = \frac{b-1}{\log b} \tanh\left(\frac{\log b}{4n}\right).
\end{aligned}$$

□

*Remark 2.11* (i) Defining $\Phi(z) = (\tanh z)/z$ and $\Phi(0) = 1$, the minimal distance given by the right-hand side in (2.7) is

$$\frac{b-1}{\log b} \tanh\left(\frac{\log b}{4n}\right) = \frac{b-1}{4n} \Phi\left(\frac{\log b}{4n}\right).$$

The function $\Phi$ is analytic, strictly decreasing on $\mathbb{R}^+$, and $\Phi(z) = 1 - \frac{1}{3}z^2 + \mathcal{O}(z^4)$. Hence, for every data set $X$ of size $n$,

$$d_K(\langle P_X\rangle_b, \mathbb{B}_b) \ge \frac{b-1}{4n}\left(1 - \frac{\log^2 b}{48n^2} + \mathcal{O}\left(\frac{\log^4 b}{n^4}\right)\right) \quad \text{as } n \to \infty,$$

so the distance between $\mathbb{B}_b$ and any $n$-point data set is at least $\mathcal{O}(1/n)$.

(ii) If, more generally, $P \in \mathcal{P}(\mathbb{R})$ is any probability measure with $\#\operatorname{supp} P \le n$ (i.e., $P$ is purely atomic with at most $n$ atoms), then $d_K(P, \mathbb{B}_b)$ can be smaller than the right-hand side in (2.7). However, the universal estimate, differing from (2.7) by merely one symbol,

$$d_K(P, \mathbb{B}_b) \ge \frac{b-1}{4n}\Phi\left(\frac{\log b}{4}\right)$$

holds, with equality for a unique $P$ having exactly $n$ atoms in $(1, b)$; see [3] for details.

Finally, to develop the concept of scale-distortion for finite data sets in the next section, the following proposition records a useful relationship between the Kantorovich metric and the $\ell_1$-norm $\|\cdot\|_1$ on $\mathbb{R}^n$,

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|.$$

For the data set $X = \{x_1, \ldots, x_n\}$, let $x_{1,n} \le x_{2,n} \le \cdots \le x_{n,n}$ be the order statistics of $X$; e.g., $x_{1,n} = \min_{1\le i\le n} x_i$ and $x_{n,n} = \max_{1\le i\le n} x_i$.

**Proposition 2.12** *Let $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_n\}$ be real data sets. Then*

$$d_K(P_X, P_Y) = \frac{1}{n}\left\|(x_{1,n}, \ldots, x_{n,n}) - (y_{1,n}, \ldots, y_{n,n})\right\|_1.$$

*Proof* Without loss of generality, assume that $x_1 \le x_2 \le \cdots \le x_n$ and $y_1 \le y_2 \le \cdots \le y_n$, so $x_i = x_{i,n}$ and $y_i = y_{i,n}$ for all $i = 1, \ldots, n$. Let $F_{P_X}$ and $F_{P_Y}$ be the d.f.'s of $P_X$ and $P_Y$, respectively, so that

$$F_{P_X}(t) = P_X\big((-\infty, t]\big) = \frac{1}{n}\#\{i \le n : x_i \le t\} \quad \text{for all } t \in \mathbb{R},$$

and similarly for $F_{P_Y}$. Note that

$$F_{P_X}^{-1}(t) = x_i \quad \text{and} \quad F_{P_Y}^{-1}(t) = y_i \quad \text{for all } t \in \left[\frac{i-1}{n}, \frac{i}{n}\right).$$

Consequently, by (2.1)

$$d_K(P_X, P_Y) = \int_0^1 |F_{P_X}^{-1}(t) - F_{P_Y}^{-1}(t)|\,dt = \sum_{i=1}^{n}\left(\frac{i}{n} - \frac{i-1}{n}\right)|x_i - y_i|$$

$$\begin{aligned}&= \frac{1}{n}\sum_{i=1}^{n}|x_i - y_i| \\ &= \frac{1}{n}\big\|(x_1,\ldots,x_n) - (y_1,\ldots,y_n)\big\|_1. \qquad \square\end{aligned}$$

*Example 2.13* For $b = 10$, the unique 2-point and 3-point data sets closest to $\mathbb{B}_{10}$ in the Kantorovich metric are $\{10^{1/4}, 10^{3/4}\}$ and $\{10^{1/6}, 10^{1/2}, 10^{5/6}\}$, respectively. Moreover, for example, every other 3-point data set is at a distance from $\mathbb{B}_{10}$ strictly larger than

$$\frac{9}{\log 10}\left(\frac{10^{1/6}-1}{10^{1/6}+1}\right) \approx 0.741.$$

*Remark 2.14* Even when the data sets $X$ and $Y$ are of different size, say $m$ and $n$, respectively, Proposition 2.12 can be applied by creating new data sets $\widehat{X}$ and $\widehat{Y}$ with $P_{\widehat{X}} = P_X$ and $P_{\widehat{Y}} = P_Y$. The points in $\widehat{X}$ are those in $X$ repeated $n/\gcd(m,n)$ times, and the points in $\widehat{Y}$ are those in $Y$ repeated $m/\gcd(m,n)$ times.

## 3 Scale-Distortion

With the tools developed in the previous section, the scale-distortion of probability measures and data sets will now be defined and analyzed. Recall that the base $b \in \mathbb{N} \setminus \{1\}$ is fixed.

**Definition 3.1** For any Borel probability measure $P$ on $\mathbb{R}^+$, let $\langle P\rangle_b$ denote the probability measure on $[1,b)$ induced via the (base $b$) mantissa function $x \mapsto \langle x\rangle_b$, i.e., the distribution function of $\langle P\rangle_b$ is given by

$$F_{\langle P\rangle_b}(t) = P(\{u : \langle u\rangle_b \le t\}) \quad \text{for all } t \in [1,b).$$

Note that this notation is consistent with the earlier use of $\langle P_X\rangle_b$.

*Example 3.2* If $P \in \mathcal{P}[1,b)$, e.g., $P = \mathbb{B}_b$ or $P$ uniform on $[1,b)$, then $\langle P\rangle_b = P$.

*Example 3.3* Let $P$ be uniform on (0, 1]. Then $\langle P\rangle_b$ is the Borel probability measure on $[1,b)$ with d.f. given by

$$\begin{aligned}F_{\langle P\rangle_b}(t) = P(\{u : \langle u\rangle_b \le t\}) = P\left(\bigcup_{n=1}^{\infty}[b^{-n}, tb^{-n}]\right) \\ = \sum_{n=1}^{\infty}(t-1)b^{-n} = \frac{t-1}{b-1}.\end{aligned}$$

Hence, $\langle P\rangle_b$ is uniform on $[1,b)$. This could be seen directly and without any computation by observing that the map $T : x \mapsto (\langle x\rangle_b - 1)/(b-1)$ on $(0,1]$ has countably many full (that is, onto) linear branches and hence preserves Lebesgue measure on $(0, 1]$, i.e., the uniform distribution $P$; see [7].

**Definition 3.4** For any Borel probability measure $P$ on $\mathbb{R}^+$ and any real number $s > 0$, the *scaling* (or *dilation*) of $P$ by $s$, denoted by $sP$, is the probability measure on $\mathbb{R}^+$ induced via the scaling $x \mapsto sx$, i.e.,

$$F_{sP}(t) = (sP)((0, t]) = P((0, t/s]) = F_P(t/s) \quad \text{for all } t > 0.$$

*Example 3.5* If $P$ is uniform on $(0, 1]$ then $sP$ is uniform on $(0, s]$. If $X = \{x_1, \dots, x_n\}$, then scaling by $s$ gives the scaled data set $sX = \{sx_1, \dots, sx_n\}$ so that $sP_X = P_{sX}$.

**Definition 3.6** Given a probability measure $P$ on $\mathbb{R}^+$ and $s > 0$, the (base $b$) *scale-distortion* of $P$ by $s$ is defined by

$$D_S(s; P) = d_K(\langle P\rangle_b, \langle sP\rangle_b).$$

The function $D_S(\cdot; P)$ quantifies how much $P$ changes under scaling. A few simple properties of this function are contained in the following lemma.

**Lemma 3.7** *Let $P$ be a Borel probability measure on $\mathbb{R}^+$, and $b \in \mathbb{N}\backslash\{1\}$. Then, for every $s \in \mathbb{R}^+$:*

(i) $D_S(sb^k; P) = D_S(s; P)$ *for all* $k \in \mathbb{Z}$;
(ii) $0 \le D_S(s; P) < b - 1$;
(iii) *The function* $D_S(\cdot; P)$ *is right-continuous*, $\lim_{\sigma\uparrow s} D_S(\sigma; P)$ *exists, and*

$$\left| D_S(s; P) - \lim_{\sigma\uparrow s} D_S(\sigma; P) \right| \le (b-1)P(\{b^k/s : k \in \mathbb{Z}\}).$$

*In particular, $D(\cdot; P)$ has at most countably many discontinuities all of which are jumps, and is continuous at $s$ whenever $P(\{b^k/s : k \in \mathbb{Z}\}) = 0$.*

*Proof* Note first that, for every $s \in \mathbb{R}^+$,

$$F_{\langle sP\rangle_b}(t) = \sum_{k\in\mathbb{Z}} (F_P(b^k t/s) - F_P(b^k/s)) + P(\{b^k/s : k \in \mathbb{Z}\}) \quad \text{for all } t \in [1, b). \tag{3.1}$$

(i) Replacing $s$ by $sb^k$ with any $k \in \mathbb{Z}$ leaves the right-hand side of (3.1) unchanged. Hence $\langle sb^k P\rangle_b = \langle sP\rangle_b$, and so $D_S(sb^k; P) = D_S(s; P)$.

(ii) Since $\langle P\rangle_b$ and $\langle sP\rangle_b$ are both elements of $\mathcal{P}[1, b)$,

$$0 \le D_S(s; P) = \int_1^b |F_{\langle P\rangle_b}(t) - F_{\langle sP\rangle_b}(t)|\, dt < \int_1^b 1\, dt = b - 1,$$

unless $|F_{\langle P\rangle_b}(t) - F_{\langle sP\rangle_b}(t)| = 1$ for almost all $t \in [1, b)$, and thus $F_{\langle P\rangle_b}(t) \in \{0, 1\}$. In the latter case, $\langle P\rangle_b = \delta_a$ for some $a \in [1, b)$. A direct computation shows that

$$D_S(s; \delta_1) = s - 1 < b - 1 \quad \text{for all } s \in [1, b),$$

and, for all $a \neq 1$,

$$D_S(s; \delta_a) = \begin{cases} a(s-1) & \text{if } 1 \le s < \frac{b}{a}, \\ a - \frac{a}{b}s & \text{if } \frac{b}{a} \le s < b, \end{cases}$$

so that $D_S(s; \delta_a) \le \max\{b-a, a-1\} < b-1$. In either case, therefore, $D_S(s; P) < b-1$, by virtue of (i).

(iii) It follows from the right-continuity of $F_P$ and (3.1) that

$$\begin{aligned} \lim_{\sigma \uparrow s} F_{\langle \sigma P \rangle_b}(t) &= \sum_{k \in \mathbb{Z}} (F_P(b^k t/s) - F_P(b^k/s)) \\ &= F_{\langle s P \rangle_b}(t) - P(\{b^k/s : k \in \mathbb{Z}\}) \quad \text{for all } t \in [1, b), \end{aligned} \tag{3.2}$$

and also

$$\begin{aligned} \lim_{\sigma \downarrow s} F_{\langle \sigma P \rangle_b}(t) &= \sum_{k \in \mathbb{Z}} (F_P(b^k t/s) - P(\{b^k t/s\}) - F_P(b^k/s) + P(\{b^k/s\})) \\ &= F_{\langle s P \rangle_b}(t) - P(\{b^k t/s : k \in \mathbb{Z}\}) \quad \text{for all } t \in [1, b). \end{aligned}$$

Consequently,

$$\lim_{\sigma \downarrow s} F_{\langle \sigma P \rangle_b}(t) = F_{\langle s P \rangle_b}(t) \quad \text{for all but countably many } t. \tag{3.3}$$

Therefore

$$\begin{aligned} &\limsup_{\sigma \downarrow s} \left| D_S(\sigma; P) - D_S(s; P) \right| \\ &\quad = \limsup_{\sigma \downarrow s} \left| d_K(\langle P \rangle_b, \langle \sigma P \rangle_b) - d_K(\langle P \rangle_b, \langle s P \rangle_b) \right| \\ &\quad \le \limsup_{\sigma \downarrow s} d_K(\langle \sigma P \rangle_b, \langle s P \rangle_b) \\ &\quad = \limsup_{\sigma \downarrow s} \int_1^b \left| F_{\langle \sigma P \rangle_b}(t) - F_{\langle s P \rangle_b}(t) \right| dt = 0, \end{aligned}$$

where the last equality follows from (3.3) and the Dominated Convergence Theorem.

Hence $\lim_{\sigma \downarrow s} D_S(\sigma; P) = D_S(s; P)$, i.e., the scale-distortion function is right-continuous. By (3.2),

$$\begin{aligned} \lim_{\sigma \uparrow s} d_K(\langle P \rangle_b, \langle \sigma P \rangle_b) &= \lim_{\sigma \uparrow s} \int_1^b \left| F_{\langle P \rangle_b}(t) - F_{\langle \sigma P \rangle_b}(t) \right| dt \\ &= \int_1^b \left| F_{\langle P \rangle_b}(t) - F_{\langle s P \rangle_b}(t) + P(\{b^k/s : k \in \mathbb{Z}\}) \right| dt, \end{aligned}$$

and so $\lim_{\sigma \uparrow s} D_S(\sigma; P)$ also exists. Moreover,

$$\begin{aligned} |D_S(s; P) - \lim_{\sigma \uparrow s} D_S(\sigma; P)| &\le \int_1^b |P(\{b^k/s : k \in \mathbb{Z}\})| \, dt \\ &= (b-1) P(\{b^k/s : k \in \mathbb{Z}\}). \end{aligned}$$

Thus if $P(\{b^k/s : k \in \mathbb{Z}\}) = 0$ then the two one-sided limits coincide, and $D_S(\cdot\,; P)$ is continuous at $s$. Observing that $P(\{b^k/s : k \in \mathbb{Z}\}) \neq 0$ for at most countably many $s$ completes the proof. □

*Example 3.8* Let $P$ be uniform on $[1, b)$. Then $\langle P\rangle_b = P$, and a short computation shows that

$$D_S(s; P) = \frac{(s-1)(b-s)}{2s} \quad \text{for all } s \in [1, b).$$

Since $F_P$ is continuous, so is the scale-distortion function $D_S(\cdot\,; P)$.

*Example 3.9* The condition $P(\{b^k/s : k \in \mathbb{Z}\}) = 0$ is not necessary for the continuity of $D_S(\cdot\,; P)$ at $s$. If, for example, $P = \delta_{(b+1)/2}$, then

$$D_S(s; P) = \frac{b+1}{4b}\big((b-1)s - |(b+1)s - 2b|\big) \quad \text{for all } s \in [1, b),$$

so that $D_S(\cdot\,; P)$ is continuous everywhere, even though $P(\{b^k/s : k \in \mathbb{Z}\}) = 1$ for $s = 2b/(1+b)$. If, on the other hand, $P = \delta_{\sqrt{b}}$ then $P(\{b^k/s : k \in \mathbb{Z}\}) = 1$ for $s = \sqrt{b}$, and $D_S(\cdot\,; P)$ has a jump there, because

$$D_S(\sqrt{b}; P) - \lim_{s\uparrow\sqrt{b}} D_S(s; P) = -(\sqrt{b}-1)^2 < 0.$$

By Lemma 3.7(ii), $D_S(\cdot\,; P)$ is bounded by $b-1$. However, a maximum may not be attained, as can be seen in Example 3.9 where $D_S(s; \delta_{\sqrt{b}}) < \sqrt{b}(\sqrt{b}-1)$ for all $s \in \mathbb{R}^+$, and yet $\sup_{s\in\mathbb{R}^+} D_S(s; \delta_{\sqrt{b}}) = \sqrt{b}(\sqrt{b}-1)$. Also, if $P$ has atoms then $D_S(\cdot\,; P)$ is in general neither upper nor lower semi-continuous. Nevertheless, the supremum of $D_S(\cdot\,; P)$ provides a useful indicator of how far $P$ is from being scale-invariant.

**Definition 3.10** The (base $b$) *scale-distortion* $D_S(P)$ of a Borel probability measure $P$ on $\mathbb{R}^+$ is

$$D_S(P) = \sup_{s\in\mathbb{R}^+} D_S(s; P) = \sup_{s\in\mathbb{R}^+} d_K(\langle P\rangle_b, \langle sP\rangle_b). \tag{3.4}$$

For a data set $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^+$ the *scale-distortion* of $X$ is $D_S(X) = D_S(P_X)$.

*Example 3.11* Let $P$ be uniform on $[1, b)$. It immediately follows from Example 3.8 that $D_S(s; P)$ is maximal for $s \in \{b^{k+1/2} : k \in \mathbb{Z}\}$, and $D_S(P) = \frac{1}{2}(\sqrt{b}-1)^2$.

*Example 3.12* A simple computation shows that $\langle s\mathbb{B}_b\rangle_b = \mathbb{B}_b$ for all $s > 0$, and therefore $D_S(\mathbb{B}_b) = 0$. In fact, if $P$ is any Borel probability measure on $\mathbb{R}^+$ then $D_S(P) = 0$ if and only if $\langle P\rangle_b = \mathbb{B}_b$, see Theorem 3.15(iii) below.

*Example 3.13* If $P = \delta_{(b+1)/2}$ then Example 3.9 shows that $D_S(P) = \frac{1}{2}(b-1)$, and also $D_S(\delta_{\sqrt{b}}) = \sqrt{b}(\sqrt{b}-1)$. Note that $D_S(\delta_{(b+1)/2}) < D_S(\delta_{\sqrt{b}})$. In fact

$D_S(\delta_{(b+1)/2}) \le D_S(\delta_a)$ for every $a > 0$, and equality holds exactly if $a = \frac{1}{2}b^k(b+1)$ for some $k \in \mathbb{Z}$; see Theorem 3.22 below.

*Remark 3.14* Scaling defines a (continuous) action of the multiplicative group $\mathbb{R}^+$ on the space of probability measures on $\mathbb{R}^+$. Via projection onto the mantissa, i.e., via $P \mapsto \langle P\rangle_b$, scaling also defines a (discontinuous) action of $\mathbb{R}^+$ on the space of probability measures on $[1, b)$. Here, the multiplicative subgroup consisting of powers of $b$ acts as the identity. Consequently, the action of $\mathbb{R}^+$ descends to an action of the quotient group $\mathbb{R}^+/\{b^k : k \in \mathbb{Z}\}$ which, as a topological group, is isomorphic to the circle. Thus to compute the scale-distortion $D_S(P)$ of $P$ it suffices to take the supremum in (3.4) over $1 \le s < b$; the latter is also evident from Lemma 3.7(i).

The next theorem summarizes the basic properties of scale-distortion.

**Theorem 3.15** *Let $P$ be a probability measure on $\mathbb{R}^+$, and $b \in \mathbb{N}\backslash\{1\}$. Then*:

(i) $0 \le D_S(P) \le b - 1$;
(ii) $D_S(\langle P\rangle_b) = D_S(P)$;
(iii) $D_S(P) = 0$ *if and only if* $\langle P\rangle_b = \mathbb{B}_b$;
(iv) $D_S(P) = b - 1$ *if and only if* $\langle P\rangle_b = \delta_1$;
(v) *If $P$ has no atoms, and if $(P_n)$ is a sequence of probability measures on $\mathbb{R}^+$ with $P_n \to P$ weakly, then $D_S(P_n) \to D_S(P)$, i.e., $D_S$ is continuous at $P$.*

*Proof* (i) This is an obvious consequence of Lemma 3.7(ii).

(ii) This follows immediately from the fact that $\langle s\langle t\rangle_b\rangle_b = \langle st\rangle_b$ for all $s, t \in \mathbb{R}^+$.

(iii) Consider the continuous map $p : \mathbb{R}^+ \to S^1$ defined as $p(t) = e^{2\pi i \log_b t}$ and note that $p(\langle t\rangle_b) = p(t)$ as well as $p(st) = p(s)p(t) = R_{\log_b s} \circ p(t)$ for all $s, t \in \mathbb{R}^+$; here $R_\vartheta$ denotes the counter-clockwise rotation of $S^1$ by an angle $2\pi\vartheta$. Clearly, $D_S(P) = 0$ if and only if $\langle sP\rangle_b = \langle P\rangle_b$ for all $s > 0$. In this case, the probability measure $\langle P\rangle_b \circ p^{-1}$ on $S^1$ satisfies

$$\langle P\rangle_b \circ p^{-1} = \langle sP\rangle_b \circ p^{-1} = (sP) \circ p^{-1} = R_{\log_b s}(P \circ p^{-1}) = R_{\log_b s}(\langle P\rangle_b \circ p^{-1}),$$

i.e., $\langle P\rangle_b \circ p^{-1}$ is invariant under *all* rotations of $S^1$. Consequently, $\langle P\rangle_b \circ p^{-1}$ equals (normalized) Lebesgue measure on $S^1$. This in turn implies that

$$\begin{aligned} F_{\langle P\rangle_b}(t) &= \langle P\rangle_b([1,t]) = \langle P\rangle_b \circ p^{-1}(\{e^{2\pi i u} : 0 \le u \le \log_b t\}) \\ &= \log_b t \quad \text{for all } t \in [1, b). \end{aligned}$$

Hence $\langle P\rangle_b = \mathbb{B}_b$. The converse, i.e. $D_S(\mathbb{B}_b) = 0$, is now obvious.

(iv) The proof of Lemma 3.7(ii) has shown that $D_S(P) < b - 1$ for every $P \in \mathcal{P}[1, b)$ with $P \ne \delta_1$, and $D_S(\delta_1) = b - 1$. Generally, therefore, $D_S(P) = b - 1$ if and only if $\langle P\rangle_b = \delta_1$.

(v) Since $P$ has no atoms, $F_{\langle sP_n\rangle_b}(t) \to F_{\langle sP\rangle_b}(t)$ for all $t \in [1, b)$ holds uniformly in $s \in [1, b)$, as does

$$\begin{aligned} \big|D_S(s; P_n) - D_S(s; P)\big| &= \big|d_K(\langle P_n\rangle_b, \langle sP_n\rangle_b) - d_K(\langle P\rangle_b, \langle sP\rangle_b)\big| \\ &\le d_K(\langle P_n\rangle_b, \langle P\rangle_b) + d_K(\langle sP_n\rangle_b, \langle sP\rangle_b) \to 0. \end{aligned}$$

Given $\varepsilon > 0$, there exists $s \in [1, b)$ such that $D_S(s; P) \geq D_S(P) - \frac{1}{2}\varepsilon$, and, for all sufficiently large $n$,

$$D_S(P_n) \geq D_S(s; P_n) \geq D_S(s; P) - \frac{1}{2}\varepsilon \geq D_S(P) - \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, $\liminf_{n\to\infty} D_S(P_n) \geq D_S(P)$. On the other hand, $D_S(s; P_n) \leq D_S(s; P) + \varepsilon \leq D_S(P) + \varepsilon$ for all sufficiently large $n$ and all $s$, so that $D_S(P_n) \leq D_S(P) + \varepsilon$. Hence $\limsup_{n\to\infty} D_S(P_n) \leq D_S(P)$, and so $\lim_{n\to\infty} D_S(P_n) = D_S(P)$. □

**Corollary 3.16** *For every $\rho \in [0, b-1]$ there exists a Borel probability measure $P$ on $\mathbb{R}^+$ such that $D_S(P) = \rho$.*

*Proof* Let $P = \frac{\rho}{b-1}\delta_1 + (1 - \frac{\rho}{b-1})\mathbb{B}_b$. Obviously, $P \in \mathcal{P}(\mathbb{R})$ if and only if $0 \leq \rho \leq b-1$, and a short calculation confirms that $D_S(s; P) = \rho\frac{s-1}{b-1}$, and hence $D_S(P) = \rho$. □

*Remark 3.17* (i) A slight refinement of the argument proving Theorem 3.15(v) shows that $P(\{b^k : k \in \mathbb{Z}\}) = 0$ is enough to ensure that $\liminf_{n\to\infty} D_S(P_n) \geq D_S(P)$ whenever $P_n \to P$ weakly, i.e., $D_S$ is lower semi-continuous at $P$. If, however, $P(\{b^k : k \in \mathbb{Z}\}) > 0$ then this is no longer true in general. For a simple example consider $P_n = \frac{1}{2}(\delta_{n/(n+1)} + \delta_1)$ for which $P_n \to \delta_1$ weakly, yet $D_S(P_n) < \frac{1}{2}(b-1)$ for all $n$. At the time of writing the authors do not know of any probability measure $P$ on $\mathbb{R}^+$ for which $D_S$ is not *upper* semi-continuous at $P$.

(ii) Convex combinations of $\delta_1$ and $\mathbb{B}_b$, as used in the proof of Corollary 3.16, are exactly the probability measures on $[1, b)$ identified as *base-invariant* in [10].

*Example 3.18* Consider the space of two-point (ordered) data sets in $[1, 10)$, i.e. $\{(x_1, x_2) : 1 \leq x_1 \leq x_2 < 10\}$. Scaling moves a point $(x_1, x_2)$ along the straight line connecting it with the origin until either the first coordinate reaches 1 or the second coordinate reaches 10. The boundary points $(a, 10)$ and $(1, a)$ are identified. Therefore, it is easy to see that the trajectory under scaling of a two-point set consists of at most two line segments. For $X = \{2, 4\}$ and $b = 10$ one segment goes from $(1, 2)$ to $(5, 10)$ and the other segment goes from $(1, 5)$ to $(2, 10)$; see Fig. 2. The point on the trajectory of $(2, 4)$ most distant from the latter (w.r.t. the $\ell_1$-metric on $\mathbb{R}^2$) clearly is $(5, 10)$, corresponding to $s = \frac{5}{2}$, and therefore $D_S(X) = \lim_{s\uparrow\frac{5}{2}} D_S(s; P_X) = \frac{1}{2}\|(2, 4) - (5, 10)\|_1 = \frac{9}{2}$, by Proposition 2.12. Also indicated in Fig. 2 by means of a dashed line is the trajectory corresponding to the scaling of the data set $X^* = \left\{\frac{1+\sqrt{10}}{2}, \frac{\sqrt{10}+10}{2}\right\}$, which is the unique two-point set in $[1, 10)$ with minimal (base 10) scale-distortion, see Theorem 3.22 below.

The next theorem provides a characterization of Benford sequences in terms of limits of the scale-distortions of the first $n$ points in the sequence. In principle, this yields a test of whether data sets are Benford or not. Since conformance to the logarithmic Benford distribution is now widely used for fraud detection and as a diag-

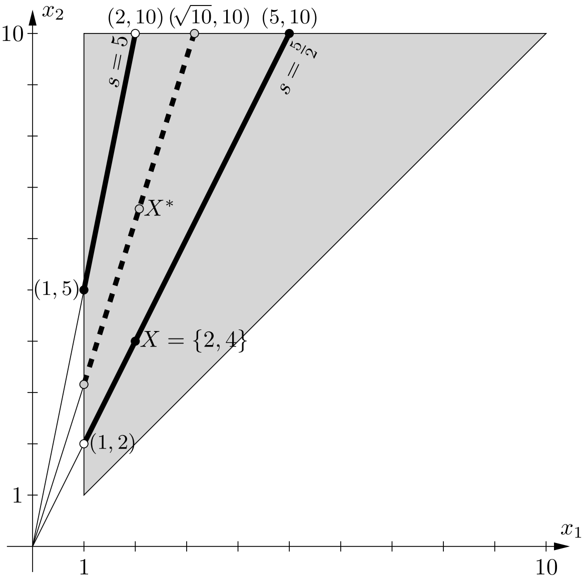

**Fig. 2** The trajectory of $X=\{2,4\}$ under scaling consists of two line segments (*solid line*). The data set $X^*=\left\{\frac{1+\sqrt{10}}{2},\frac{\sqrt{10}+10}{2}\right\}$ has minimal (base 10) scale-distortion and its scaling trajectory consists of one segment only (*dashed line*), see Examples 3.18 and 3.23

nostic test for mathematical models, the scale-distortion characterization may prove to be a useful alternative in practical applications.

**Theorem 3.19** *Let $(x_n)$ be a sequence in $\mathbb{R}^+$ and $X_n=\{x_1,\dots,x_n\}$. Then $(x_n)$ is b-Benford if and only if $D_S(X_n)\to 0$ as $n\to\infty$.*

The next lemma will be used in the proof of this theorem.

**Lemma 3.20** *Let $P$ be a probability measure on $\mathbb{R}^+$ with $\langle P\rangle_b \neq \mathbb{B}_b$. Then there exists $s^*\in[1,b)$ such that*

(i) $\langle s^*P\rangle_b \neq \langle P\rangle_b$ *and*
(ii) $P(\{b^k/s^*:k\in\mathbb{Z}\})=0$.

*Proof of Lemma 3.20* The first statement is immediate from Lemma 3.15(iii), and in case $P$ has no atoms the overall statement is obvious. Assume, therefore, that $P$ has an atom. Then $P(\{a\})=\varepsilon>0$ for some $a\in\mathbb{R}^+$, and so $\langle sP\rangle_b(\{\langle sa\rangle_b\})\geq\varepsilon$ for all $s$. This implies that $\langle sP\rangle_b\neq\langle P\rangle_b$ for those $s$ for which $\langle P\rangle_b(\{\langle sa\rangle_b\})<\varepsilon$, that is,

$$\langle sP\rangle_b\neq\langle P\rangle_b \quad \text{for all but a finite number of } s \text{ in } [1,b) \tag{3.5}$$

since $P$ is a probability measure. Furthermore,

$$P(\{b^k/s:k\in\mathbb{Z}\})=0 \quad \text{for all but a countable number of } s \text{ in } [1,b). \tag{3.6}$$

By (3.5) and (3.6) properties (i) and (ii) hold simultaneously for all $s$ from an appropriate set $S\subset[1,b)$, where $[1,b)\backslash S$ is countable. □

*Proof of Theorem 3.19* Assume first that $(x_n)$ is $b$-Benford. By Proposition 2.3 this means that $\langle P_{X_n}\rangle_b \to \mathbb{B}_b$ weakly. Since $\mathbb{B}_b$ does not have atoms,

$$D_S(X_n) = D_S(\langle P_{X_n}\rangle_b) \to D_S(\mathbb{B}_b) = 0,$$

by Proposition 3.15(v) and Example 3.12.

Conversely, suppose that $(x_n)$ is not $b$-Benford. Since $\langle P_{X_n}\rangle_b \in \mathcal{P}[1,b)$, the family $\{\langle P_{X_n}\rangle_b : n \in \mathbb{N}\}$ is tight and so contains a convergent subsequence [5, Theorem 29.3]; let $P_n = \langle P_{X_n}\rangle_b$ and assume without loss of generality that $P_n \to P$ for some probability measure $P \neq \mathbb{B}_b$. By Lemma 3.20 there exists $s^* \in [1,b)$ and $\delta > 0$ such that $d_K(P, \langle s^*P\rangle_b \geq \delta$ and $P(\{b^k/s^* : k \in \mathbb{Z}\}) = 0$. It follows from (3.1) and the definition of weak convergence that $F_{\langle s^*P_n\rangle_b}(t) \to F_{\langle s^*P\rangle_b}(t)$ for almost all $t \in [1,b)$, hence $d_K(\langle s^*P_n\rangle_b, \langle s^*P\rangle_b) \to 0$. Since $d_K$ metrizes weak convergence,

$$\begin{aligned} D_S(X_n) &\geq d_K(P_n, \langle s^*P_n\rangle_b) \geq d_K(P, \langle s^*P\rangle_b) - d_K(P_n, P) - d_K(\langle s^*P_n\rangle_b, \langle s^*P\rangle_b) \\ &\to d_K(\langle P\rangle_b, \langle s^*P\rangle_b) > 0. \end{aligned}$$

Thus $\limsup_{n\to\infty} D_S(X_n) \geq d_K(P, \langle s^*P\rangle_b) > 0$. □

Theorem 3.19 has the following natural analogue in a statistical setting.

**Theorem 3.21** *Suppose $X_1, X_2, \ldots$ are independent, identically distributed random variables on $\mathbb{R}^+$ with common distribution $P$. Then*

(i) $\langle P\rangle_b = \mathbb{B}_b$ *if and only if* $D_S(\{X_1,\ldots,X_n\}) \to 0$ *almost surely as* $n \to \infty$;
(ii) $\langle P\rangle_b \neq \mathbb{B}_b$ *if and only if* $\limsup_{n\to\infty} D_S(\{X_1,\ldots,X_n\}) > 0$ *almost surely.*

*Proof* For each $n \in \mathbb{N}$ let $F_n$ denote the empirical distribution function for $X_1,\ldots,X_n$, i.e., $F_n(t) = P_n((-\infty,t])$, where $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$. By the Glivenko-Cantelli Theorem [5, Theorem 20.6], $F_n$ converges to $F_P$ uniformly almost surely, so, almost surely, $P_n \to P$ weakly. Conclusions (i) and (ii) then follow directly from Theorem 3.19. □

The next result is the main scale-distortion inequality in this article. It identifies, for every positive integer $n$, the unique data set of $n$ points that is least distorted by change of scale, e.g., by change of monetary or physical units, and it identifies the minimal scale-distortion attained by any $n$-point set.

**Theorem 3.22** *Let $n \in \mathbb{N}$ and let $X = \{x_1,\ldots,x_n\} \subset \mathbb{R}^+$ be an $n$-point data set. Then $D_S(X) \geq (b-1)/(2n)$, and equality holds if and only if*

$$\{\langle x_1\rangle_b,\ldots,\langle x_n\rangle_b\} = \left\{\frac{1+b^{1/n}}{2} b^{(i-1)/n} : i = 1,\ldots,n\right\}. \tag{3.7}$$

*Proof* Let $y_i = \langle x_i\rangle_b$ for $i = 1,\ldots,n$, and assume without loss of generality that $1 \leq y_1 \leq \cdots \leq y_n < b$. Hence $\{y_1,\ldots,y_n\}$ is an $n$-point ordered data set in $[1,b)$. Identify the space of all such data sets with the subset of $\mathbb{R}^n$ given by $\{y \in \mathbb{R}^n : 1 \leq y_1 \leq \cdots \leq$

$y_n < b\}$. The scaling trajectory of $y$, i.e. the set $\{\langle sy\rangle_b = (\langle sy_1\rangle_b, \dots, \langle sy_n\rangle_b) : s \in [1, b)\}$, is a union of at most $n$ line segments. To see this, consider the scaling of $y$ by increasing $s$, beginning with $s = 1$. The resulting line will first reach the boundary for $s = b/y_n$, that is, when the $n$th coordinate reaches $b$. The value $b$ is then replaced by 1, which becomes the new first entry of the data set. The vector representation is

$$\left(\frac{b}{y_n}\right)(y_1, y_2, \dots, y_n) = \left(1, \frac{b}{y_n}y_1, \dots, \frac{b}{y_n}y_{n-1}\right),$$

as the other components are shifted one place to the right. Then the scaling continues with increasing $s$ until the rightmost component reaches $b$, etc. Each time the rightmost coordinate reaches $b$, there is a break. The trajectory resumes with a first coordinate equal to 1 and the others shifted to the right by one place. The breaks occur for values $s = b/y_i$ and so there are $n$ breaks in the trajectory of $y$ as $s$ increases from 1 to $b$. When $s = b$ the trajectory closes at the starting point $y$.

The trajectory of $y$ can also be characterized by the $n$-tuple of ratios $(r_1, r_2, \dots, r_n)$ where $r_i = y_i/y_{i-1}$for $i = 2, \dots, n$ and $r_1 = by_1/y_n$. Clearly, all the ratios $r_i$ are numbers in $[1, b]$, and they satisfy $\prod_{i=1}^n r_i = b$. Any $(r_1, r_2, \dots, r_n)$ with these properties is associated to a scaling trajectory, and two $n$-tuples of ratios describe the same trajectory when they are cyclic permutations of each other. Given $y$, assume without loss of generality that $r_1 \geq r_i$ for all $i = 1, \dots, n$. The scaling trajectory of $y$ contains the two points

$$\eta_l = (1, r_2, r_2r_3, \dots, r_2r_3\cdots r_n) = \left(1, \frac{y_2}{y_1}, \frac{y_3}{y_1}, \dots, \frac{y_n}{y_1}\right)$$

and

$$\eta_u = (r_1, r_1r_2, r_1r_2r_3, \dots, r_1r_2r_3\cdots r_n) = \left(b\frac{y_1}{y_n}, b\frac{y_2}{y_n}, b\frac{y_3}{y_n}, \dots, b\right)$$

as endpoints of one of its segments. From

$$\begin{aligned}\|\eta_u - \eta_l\|_1 &= r_1 - 1 + r_1r_2 - r_2 + r_1r_2r_3 - r_2r_3 + \cdots + r_1r_2r_3\cdots r_n - r_2r_3\cdots r_n \\ &= b - 1 + r_1 - r_2 + r_2(r_1 - r_3) + \cdots + r_2r_3\cdots r_{n-1}(r_1 - r_n) \\ &\geq b - 1, \qquad (3.8)\end{aligned}$$

it follows that the trajectory of $y$ contains a segment of $\ell_1$-length at least $b - 1$. Since $\|\eta_u - y\|_1 + \|\eta_l - y\|_1 \geq \|\eta_u - \eta_l\|_1 \geq b - 1$, one of the points $\eta_u, \eta_l$ has $\ell_1$-distance no less than $\frac{1}{2}(b - 1)$ from $y$ so that, by Proposition 2.12,

$$D_S(Y) = \sup_{s\in[1,b)} d_K(\langle P_Y\rangle_b, \langle sP_Y\rangle_b) = \frac{1}{n}\sup_{s\in[1,b)} \|\langle y\rangle_b - \langle sy\rangle_b\|_1 \geq \frac{b-1}{2n}.$$

Moreover, since $r_1 \geq r_i$ for $i = 1, \dots, n$, (3.8) implies that the latter inequality is strict unless $r_1 = r_2 = \cdots = r_n$ and hence $r_i = b^{1/n}$ for all $i$. In this case, the trajectory of

$y$ consists of a single segment whose midpoint

$$y^* = \frac{\eta_u + \eta_l}{2} = \frac{1 + b^{1/n}}{2}(1, b^{1/n}, \ldots, b^{(n-1)/n})$$

satisfies $d_K(\langle P_{Y^*}\rangle_b, \langle sP_{Y^*}\rangle_b) \le (b-1)/(2n)$ for all $s > 0$, so that $D_S(Y^*) = (b-1)/(2n)$. □

*Example 3.23* The $n$-point data set $X^* \subset [1, b)$ with minimal scale-distortion according to (3.7) is *not* identical to the data set $X \subset [1, b)$ that minimizes $d_K(P_X, \mathbb{B}_b)$, as given by Corollary 2.10. However, both data sets are geometric progressions with ratio $b^{1/n}$, and $X^*$ is a scaled version of $X$, namely, $X^* = sX$ with $s = \frac{1}{2}(b^{1/(2n)} + b^{-1/(2n)}) = \cosh(\frac{\log b}{2n})$.

For $b = 10$, $n = 2$ the data set with minimal scale-distortion is $X^* = \{\frac{1+\sqrt{10}}{2}, \frac{\sqrt{10}+10}{2}\} \approx \{2.08, 6.58\}$. Figure 2 shows that the scaling trajectory of $X^*$ is a single segment with midpoint $(x_1^*, x_2^*)$; this segment lies between the two segments of the trajectory of $X = \{2, 4\}$. Recall from Example 2.13 that the 2-point data set closest to $\mathbb{B}_{10}$ in the Kantorovich metric is $\{10^{1/4}, 10^{3/4}\} \approx \{1.78, 5.62\}$.

## References

1. Allaart, P.C.: An invariant-sum characterization of Benford's law. J. Appl. Probab. **34**, 288–291 (1997)
2. Benford, F.: The law of anomalous numbers. Proc. Am. Philos. Soc. **78**, 551–572 (1938)
3. Berger, A., Morrison, K.E.: Best finite Kantorovich approximations (in preparation)
4. Bickel, P.J., Doksum, K.A.: Mathematical Statistics. Holden-Day, San Francisco (1976)
5. Billingsley, P.: Probability and Measure, 3rd edn. Wiley, New York (1995)
6. Chung, K.L.: A Course in Probability Theory, 2nd edn. Academic, New York (1974)
7. Dajani, K., Kraaikamp, C.: Carus Mathematical Monographs. Ergodic Theory of Numbers, vol. 29. Mathematical Association of America, Washington (2002)
8. Dudley, R.M.: Real Analysis and Probability. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove (1989)
9. Gibbs, A., Su, F.E.: On choosing and bounding probability metrics. Int. Stat. Rev. **70**, 419–435 (2002)
10. Hill, T.P.: Base-invariance implies Benford's law. Proc. Am. Math. Soc. **123**, 887–895 (1995)
11. Hill, T.P.: A statistical derivation of the significant-digit law. Stat. Sci. **10**, 354–363 (1995)
12. Knuth, D.E.: The Art of Computer Programming, volume 2: Seminumerical Algorithms, 2nd edn. Addison–Wesley, Reading (1981)
13. Newcomb, S.: Note on the frequency of use of the different digits in natural numbers. Am. J. Math. **4**, 39–40 (1881)
14. Nigrini, M.: A taxpayer compliance application of Benford's law. J. Am. Tax. Assoc. **1**, 72–91 (1996)