ON CHOOSING AND BOUNDING PROBABILITY METRICS

ALISON L. GIBBS AND FRANCIS EDWARD SU

Manuscript version January 2002

Abstract. When studying convergence of measures, an important issue is the choice of probability metric. We provide a summary and some new results concerning bounds among some important probability metrics/distances that are used by statisticians and probabilists. Knowledge of other metrics can provide a means of deriving bounds for another one in an applied problem. Considering other metrics can also provide alternate insights. We also give examples that show that rates of convergence can strongly depend on the metric chosen. Careful consideration is necessary when choosing a metric.

Abrégé. Le choix de métrique de probabilité est une décision très importante lorsqu'on étudie la convergence des mesures. Nous vous fournissons avec un sommaire de plusieurs métriques/distances de probabilité couramment utilisées par des statisticiens(nes) at par des probabilistes, ainsi que certains nouveaux résultats qui se rapportent à leurs bornes. Avoir connaissance d'autres métriques peut vous fournir avec un moyen de dériver des bornes pour une autre métrique dans un problème appliqué. Le fait de prendre en considération plusieurs métriques vous permettra d'approcher des problèmes d'une manière différente. Ainsi, nous vous démontrons que les taux de convergence peuvent dépendre de façon importante sur votre choix de métrique. Il est donc important de tout considérer lorsqu'on doit choisir une métrique.

1. Introduction

Determining whether a sequence of probability measures converges is a common task for a statistician or probabilist. In many applications it is important also to *quantify* that convergence in terms of some *probability metric*; hard numbers can then be interpreted by the metric's meaning, and one can proceed to ask qualitative questions about the nature of that convergence.

Key words and phrases. discrepancy, Hellinger distance, probability metrics, Prokhorov metric, relative entropy, rates of convergence, Wasserstein distance.

First author supported in part by an NSERC postdoctoral fellowship. Second author acknowledges the hospitality of the Cornell School of Operations Research during a sabbatical in which this was completed. The authors thank Jeff Rosenthal and Persi Diaconis for their encouragement of this project and Neal Madras for helpful discussions.

Abbreviation	Metric
D	Discrepancy
H	Hellinger distance
I	Relative entropy (or Kullback-Leibler divergence)
K	Kolmogorov (or Uniform) metric
L	Lévy metric
P	Prokhorov metric
S	Separation distance
TV	Total variation distance
W	Wasserstein (or Kantorovich) metric
χ^2	χ^2 distance

Table 1. Abbreviations for metrics used in Figure 1.

There are a host of metrics available to quantify the distance between probability measures; some are not even metrics in the strict sense of the word, but are simply notions of "distance" that have proven useful to consider. How does one choose among all these metrics? Issues that can affect a metric's desirability include whether it has an interpretation applicable to the problem at hand, important theoretical properties, or useful bounding techniques.

Moreover, even after a metric is chosen, it can still be useful to familiarize oneself with other metrics, especially if one also considers the relationships among them. One reason is that bounding techniques for one metric can be exploited to yield bounds for the desired metric. Alternatively, analysis of a problem using several different metrics can provide complementary insights.

The purpose of this paper is to review some of the most important metrics on probability measures and the relationships among them. This project arose out of the authors' frustrations in discovering that, while encyclopedic accounts of probability metrics are available (e.g., Rachev (1991)), relationships among such metrics are mostly scattered through the literature or unavailable. Hence in this review we collect in one place descriptions of, and bounds between, ten important probability metrics. We focus on these metrics because they are either well-known, commonly used, or admit practical bounding techniques. We limit ourselves to metrics between probability measures (*simple metrics*) rather than the broader context of metrics between random variables (*compound metrics*).

We summarize these relationships in a handy reference diagram (Figure 1), and provide new bounds between several metrics. We also give examples to illustrate that, depending on the choice of metric, rates of convergence can differ both quantitatively and qualitatively.

This paper is organized as follows. Section 2 reviews properties of our ten chosen metrics. Section 3 contains references or proofs of the bounds in Figure 1. Some examples of their applications are described in Section 4. In

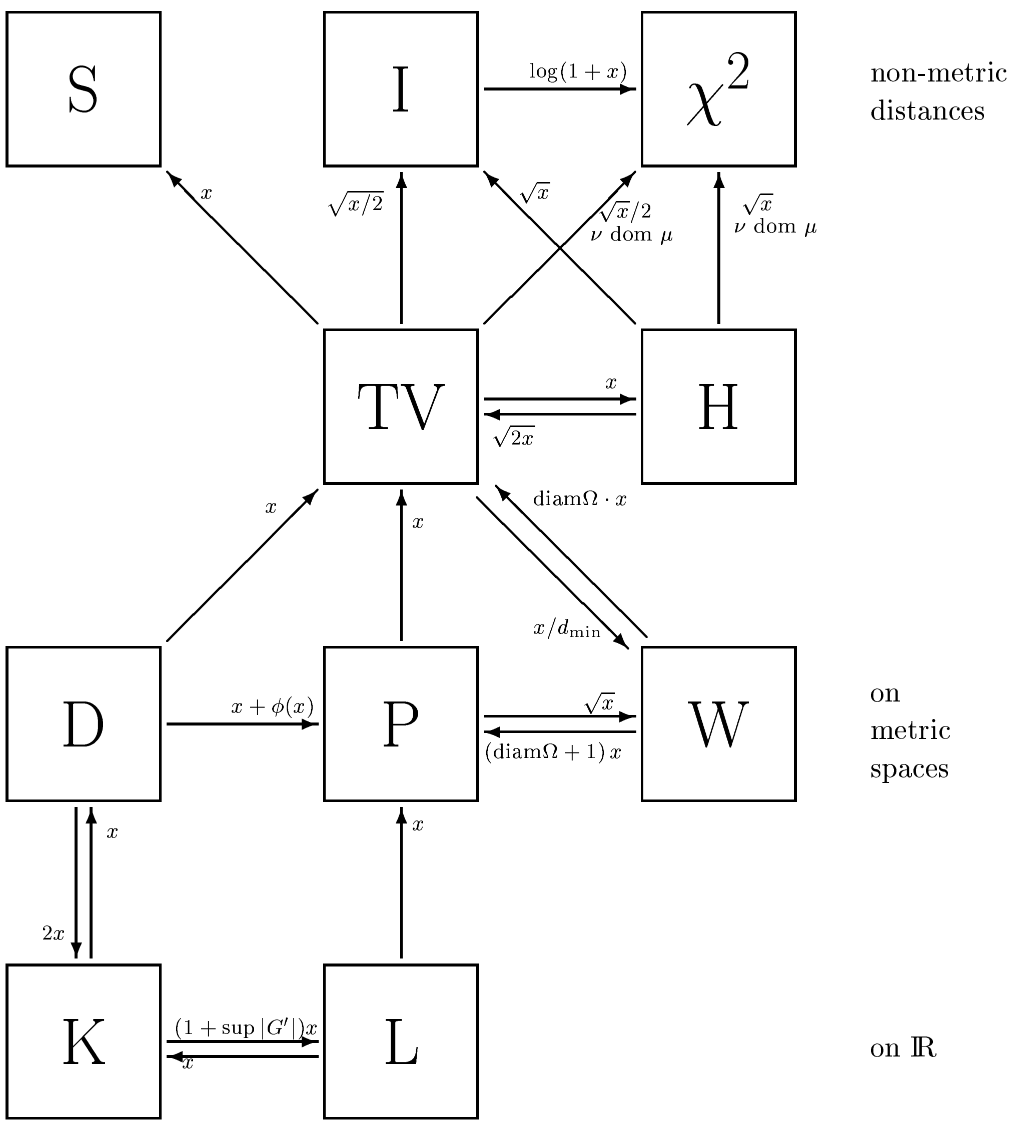

FIGURE 1. Relationships among probability metrics. A directed arrow from A to B annotated by a function $h(x)$ means that $d_A \leq h(d_B)$. The symbol diam Ω denotes the diameter of the probability space Ω; bounds involving it are only useful if Ω is bounded. For Ω finite, $d_{\min} = \inf_{x,y\in\Omega} d(x,y)$. The probability metrics take arguments μ, ν; "ν dom μ" indicates that the given bound only holds if ν dominates μ. Other notation and restrictions on applicability are discussed in Section 3.

Section 5 we give examples to show that the choice of metric can strongly affect both the rate and nature of the convergence.

2. Ten metrics on probability measures

Throughout this paper, let Ω denote a measurable space with σ-algebra $\mathcal{B}$. Let $\mathcal{M}$ be the space of all probability measures on $(\Omega, \mathcal{B})$. We consider convergence in $\mathcal{M}$ under various notions of distance, whose definitions are reviewed in this section. Some of these are not strictly metrics, but are non-negative notions of "distance" between probability distributions on Ω that have proven useful in practice. These distances are reviewed in the order given by Table 1.

In what follows, let μ, ν denote two probability measures on Ω. Let f and g denote their corresponding density functions with respect to a σ-finite dominating measure λ (for example, λ can be taken to be $(\mu+\nu)/2$). If $\Omega = \mathbb{R}$, let F, G denote their corresponding distribution functions. When needed, X, Y will denote random variables on Ω such that $\mathcal{L}(X) = \mu$ and $\mathcal{L}(Y) = \nu$. If Ω is a metric space, it will be understood to be a measurable space with the Borel σ-algebra. If Ω is a bounded metric space with metric d, let $\operatorname{diam}(\Omega) = \sup\{d(x,y) : x, y \in \Omega\}$ denote the *diameter* of Ω.

Discrepancy metric.

1. State space: Ω any metric space.
2. Definition:
$$d_D(\mu, \nu) := \sup_{\text{all closed balls } B} |\mu(B) - \nu(B)|.$$
It assumes values in $[0, 1]$.
3. The discrepancy metric recognizes the metric topology of the underlying space Ω. However, the discrepancy is scale-invariant: multiplying the metric of Ω by a positive constant does not affect the discrepancy.
4. The discrepancy metric admits Fourier bounds, which makes it useful to study convergence of random walks on groups (Diaconis 1988, p. 34).

Hellinger distance.

1. State space: Ω any measurable space.
2. Definition: if f, g are densities of the measures μ, ν with respect to a dominating measure λ,
$$d_H(\mu, \nu) := \left[\int_\Omega (\sqrt{f} - \sqrt{g})^2 \, d\lambda\right]^{1/2} = \left[2\left(1 - \int_\Omega \sqrt{fg} \, d\lambda\right)\right]^{1/2}.$$
This definition is independent of the choice of dominating measure λ. For a countable state space Ω,
$$d_H(\mu, \nu) := \left[\sum_{\omega \in \Omega} \left(\sqrt{\mu(\omega)} - \sqrt{\nu(\omega)}\right)^2\right]^{1/2}$$
(Diaconis and Zabell 1982).

3. It assumes values in $[0, \sqrt{2}]$. Some texts, e.g., LeCam (1986), introduce a factor of a square root of two in the definition of the Hellinger distance to normalize its range of possible values to $[0, 1]$. We follow Zolotarev (1983). Other sources, e.g., Borovkov (1998), Diaconis and Zabell (1982), define the Hellinger distance to be the square of d_H. (While d_H is a metric, d_H^2 is not.) An important property is that for product measures $\mu = \mu_1 \times \mu_2$, $\nu = \nu_1 \times \nu_2$ on a product space $\Omega_1 \times \Omega_2$,
$$1 - \frac{1}{2}d_H^2(\mu, \nu) = \left(1 - \frac{1}{2}d_H^2(\mu_1, \nu_1)\right)\left(1 - \frac{1}{2}d_H^2(\mu_2, \nu_2)\right)$$
(Zolotarev 1983, p. 279). Thus one can express the distance between distributions of vectors with independent components in terms of the component-wise distances. A consequence (Reiss 1989, p. 100) of the above formula is $d_H^2(\mu, \nu) \le d_H^2(\mu_1, \nu_1) + d_H^2(\mu_2, \nu_2)$.
4. The quantity $(1 - \frac{1}{2}d_H^2)$ is called the *Hellinger affinity*. Apparently Hellinger (1907) used a similar quantity in operator theory, but Kakutani (1948, p. 216) appears responsible for popularizing the Hellinger affinity and the form d_H in his investigation of infinite products of measures. Le Cam and Yang (1990) and Liese and Vajda (1987) contain further historical references.

Relative entropy (or Kullback-Leibler divergence).

1. State space: Ω any measurable space.
2. Definition: if f, g are densities of the measures μ, ν with respect to a dominating measure λ,
$$d_I(\mu, \nu) := \int_{S(\mu)} f \log(f/g)\; d\lambda,$$
where $S(\mu)$ is the support of μ on Ω. The definition is independent of the choice of dominating measure λ. For Ω a countable space,
$$d_I(\mu, \nu) := \sum_{\omega \in \Omega} \mu(\omega) \log \frac{\mu(\omega)}{\nu(\omega)}.$$
The usual convention, based on continuity arguments, is to take $0 \log \frac{0}{q} = 0$ for all real q and $p \log \frac{p}{0} = \infty$ for all real non-zero p. Hence the relative entropy assumes values in $[0, \infty]$.
3. Relative entropy is not a metric, since it is not symmetric and does not satisfy the triangle inequality. However, it has many useful properties, including additivity over marginals of product measures: if $\mu = \mu_1 \times \mu_2$, $\nu = \nu_1 \times \nu_2$ on a product space $\Omega_1 \times \Omega_2$,
$$d_I(\mu, \nu) = d_I(\mu_1, \nu_1) + d_I(\mu_2, \nu_2)$$
(Cover and Thomas 1991, Reiss 1989, p. 100).
4. Relative entropy was first defined by Kullback and Leibler (1951) as a generalization of the entropy notion of Shannon (1948). A standard reference on its properties is Cover and Thomas (1991).

Kolmogorov (or Uniform) metric.

1. State space: $\Omega = \mathbb{R}$.
2. Definition:
$$d_K(F, G) := \sup_x |F(x) - G(x)|, \quad x \in \mathbb{R}.$$
(Since μ, ν are measures on $\mathbb{R}$, it is customary to express the Kolmogorov metric as a distance between their distribution functions F, G.)
3. It assumes values in $[0, 1]$, and is invariant under all increasing one-to-one transformations of the line.
4. This metric, due to Kolmogorov (1933), is also called the uniform metric (Zolotarev 1983).

Lévy metric.

1. State space: $\Omega = \mathbb{R}$.
2. Definition:
$$d_L(F, G) := \inf\{\epsilon > 0 : G(x - \epsilon) - \epsilon \leq F(x) \leq G(x + \epsilon) + \epsilon, \forall x \in \mathbb{R}\}.$$
(Since μ, ν are measures on $\mathbb{R}$, it is customary to express the Lévy metric as a distance between their distribution functions F, G.)
3. It assumes values in $[0, 1]$. While not easy to compute, the Lévy metric does metrize weak convergence of measures on $\mathbb{R}$ (Lukacs 1975, p. 71). It is shift invariant, but not scale invariant.
4. This metric was introduced by Lévy (1925, p. 199-200).

Prokhorov (or Lévy-Prokhorov) metric.

1. State space: Ω any metric space.
2. Definition:
$$d_P(\mu, \nu) := \inf\{\epsilon > 0 : \mu(B) \leq \nu(B^\epsilon) + \epsilon \text{ for all Borel sets } B\}$$
where $B^\epsilon = \{x : \inf_{y \in B} d(x, y) \leq \epsilon\}$. It assumes values in $[0, 1]$.
3. It is possible to show that this metric is symmetric in μ, ν. See (Huber 1981, p. 27).
4. This metric was defined by Prokhorov (1956) as the analogue of the Lévy metric for more general spaces. While not easy to compute, this metric is theoretically important because it metrizes weak convergence on any separable metric space (Huber 1981, p. 28). Moreover, $d_P(\mu, \nu)$ is precisely the minimum distance "in probability" between random variables distributed according to μ, ν. This was shown by Strassen (1965) for complete separable metric spaces and extended by Dudley (1968) to arbitrary separable metric spaces.

Separation distance.

1. State space: Ω a countable space.
2. Definition:
$$d_S(\mu, \nu) := \max_i \left(1 - \frac{\mu(i)}{\nu(i)}\right).$$

3. It assumes values in $[0,1]$. However, it not a metric.
4. The separation distance was advocated by Aldous and Diaconis (1987) to study Markov chains because it admits a useful characterization in terms of strong uniform times.

Total variation distance.

1. State space: Ω any measurable space.
2. Definition:

$$d_{TV}(\mu,\nu) := \sup_{A\subset\Omega} \mid \mu(A)-\nu(A) \mid \tag{1}$$

$$= \frac{1}{2}\max_{|h|\leq 1}\left|\int h\ d\mu - \int h\ d\nu\right| \tag{2}$$

where $h:\Omega\to\mathbb{R}$ satisfies $|h(x)|\leq 1$. This metric assumes values in $[0,1]$.

For a countable state space Ω, the definition above becomes

$$d_{TV} := \frac{1}{2}\sum_{x\in\Omega} \mid \mu(x)-\nu(x) \mid$$

which is half the L^1-norm between the two measures. Some authors (for example, Tierney (1996)) define total variation distance as twice this definition.

3. Total variation distance has a coupling characterization:

$$d_{TV}(\mu,\nu) = \inf\{\Pr(X\neq Y):\ \text{r.v. } X,\ Y \text{ s.t. } \mathcal{L}(X)=\mu,\ \mathcal{L}(Y)=\nu\}$$

(Lindvall 1992, p. 19).

Wasserstein (or Kantorovich) metric.

1. State space: $\mathbb{R}$ or any metric space.
2. Definition: For $\Omega=\mathbb{R}$, if F,G are the distribution functions of μ,ν respectively, the *Kantorovich metric* is defined by

$$d_W(\mu,\nu) := \int_{-\infty}^{\infty} |F(x)-G(x)|\,dx$$

$$= \int_0^1 |F^{-1}(t)-G^{-1}(t)|\,dt.$$

Here F^{-1},G^{-1} are the inverse functions of the distribution functions F,G. For any separable metric space, this is equivalent to

$$d_W(\mu,\nu) := \sup\left\{\left|\int h\ d\mu - \int h\ d\nu\right| : \|h\|_L\leq 1\right\}, \tag{3}$$

the supremum being taken over all h satisfying the Lipschitz condition $|h(x)-h(y)|\leq d(x,y)$, where d is the metric on Ω.

3. The Wasserstein metric assumes values in $[0, \mathrm{diam}(\Omega)]$, where $\mathrm{diam}(\Omega)$ is the diameter of the metric space (Ω, d). This metric metrizes weak convergence on spaces of bounded diameter, as is evident from Theorem 2 below.
4. By the Kantorovich-Rubinstein theorem, the Kantorovich metric is equal to the *Wasserstein metric*:
$$d_W(\mu, \nu) = \inf_J \{\mathbf{E}[d(X, Y)] : \mathcal{L}(X) = \mu, \mathcal{L}(Y) = \nu\},$$
where the infimum is taken over all joint distributions J with marginals μ, ν. See Szulga (1982, Theorem 2). Dudley (1989, p. 342) traces some of the history of these metrics.

χ^2-distance.

1. State space: Ω any measurable space.
2. Definition: if f, g are densities of the measures μ, ν with respect to a dominating measure λ, and $S(\mu), S(\nu)$ are their supports on Ω,
$$d_{\chi^2}(\mu, \nu) := \int_{S(\mu) \cup S(\nu)} \frac{(f - g)^2}{g} \, d\lambda.$$
This definition is independent of the choice of dominating measure λ. This metric assumes values in $[0, \infty]$. For a countable space Ω this reduces to:
$$d_{\chi^2}(\mu, \nu) := \sum_{\omega \in S(\mu) \cup S(\nu)} \frac{(\mu(\omega) - \nu(\omega))^2}{\nu(\omega)}.$$
This distance is not symmetric in μ and ν; beware that the order of the arguments varies from author to author. We follow Csiszar (1967) and Liese and Vajda (1987) because of the remarks following equation (4) below and the natural order of arguments suggested by inequalities (10) and (11). Reiss (1989, p. 98) takes the opposite convention as well as defining the χ^2-distance as the square root of the above expression.
3. The χ^2-distance is not symmetric, and therefore not a metric. However, like the Hellinger distance and relative entropy, the χ^2-distance between product measures can be bounded in terms of the distances between their marginals. See Reiss (1989, p. 100).
4. The χ^2-distance has origins in mathematical statistics dating back to Pearson. See Liese and Vajda (1987, p. 51) for some history.

We remark that several distance notions in this section are instances of a family of distances known as *f-divergences* (Csiszar 1967). For any convex function f, one may define

$$d_f(\mu, \nu) = \sum_{\omega} \nu(\omega) f(\frac{\mu(\omega)}{\nu(\omega)}). \tag{4}$$

Then choosing $f(x) = (x-1)^2$ yields d_{χ^2}, $f(x) = x \log x$ yields d_I, $f(x) = |x-1|/2$ yields d_{TV}, and $f(x) = (\sqrt{x}-1)^2$ yields d_H^2. The family of f-divergences are studied in detail in Liese and Vajda (1987).

3. Some Relationships Among Probability Metrics

In this section we describe in detail the relationships illustrated in Figure 1. We give references for relationships known to appear elsewhere, and prove several new bounds which we state as theorems. In choosing the order of presentation, we loosely follow the diagram from bottom to top.

At the end of this section, we summarize in Theorem 6 what is known about how these metrics relate to weak convergence of measures.

The Kolmogorov and Lévy metrics on $\mathbb{R}$. For probability measures μ, ν on $\mathbb{R}$ with distribution functions F, G,

$$d_L(F,G) \leq d_K(F,G).$$

See Huber (1981, p. 34). Petrov (1995, p. 43) notes that if $G(x)$ (i.e., ν) is absolutely continuous (with respect to Lebesgue measure), then

$$d_K(F,G) \leq \left(1 + \sup_x |G'(x)|\right) d_L(F,G).$$

The Discrepancy and Kolmogorov metrics on $\mathbb{R}$. It is evident that for probability measures on $\mathbb{R}$,

$$d_K \leq d_D \leq 2\, d_K. \tag{5}$$

This follows from the regularity of Borel sets in $\mathbb{R}$ and expressing closed intervals in $\mathbb{R}$ as difference of rays.

The Prokhorov and Lévy metrics on $\mathbb{R}$. For probability measures on $\mathbb{R}$,

$$d_L \leq d_P.$$

See Huber (1981, p. 34).

The Prokhorov and Discrepancy metrics. The following theorem shows how discrepancy may be bounded by the Prokhorov metric by finding a suitable right-continuous function ϕ. For bounded Ω, $\phi(\epsilon)$ gives an upper bound on the additional ν-measure of the extended ball B^ϵ over the ball B, where $B^\epsilon = \{x : \inf_{y \in B} d(x,y) \leq \epsilon\}$. Note that this theorem also gives an upper bound for d_K through (5) above.

Theorem 1. *Let Ω be any metric space, and let ν be any probability measure satisfying*

$$\nu(B^\epsilon) \leq \nu(B) + \phi(\epsilon)$$

for all balls B and complements of balls B and some right-continuous function ϕ. Then for any other probability measure μ, if $d_P(\mu,\nu) = x$, then $d_D(\mu,\nu) \leq x + \phi(x)$.

As an example, on the circle or line, if $\nu = U$ is the uniform distribution, then $\phi(x) = 2x$ and hence

$$d_D(\mu, U) \leq 3\, d_P(\mu, U).$$

Proof. For μ, ν as above,

$$\mu(B) - \nu(B^x) \geq \mu(B) - \nu(B) - \phi(x).$$

And if $d_P(\mu, \nu) = x$, then $\mu(B) - \nu(B^{\tilde{x}}) \leq \tilde{x}$ for all $\tilde{x} > x$ and all Borel sets B. Combining with the above inequality, we see that

$$\mu(B) - \nu(B) - \phi(\tilde{x}) \leq \tilde{x}.$$

By taking the supremum over B which are balls or complements of balls, obtain

$$\sup_B (\mu(B) - \nu(B)) \leq \tilde{x} + \phi(\tilde{x}).$$

The same result may be obtained for $\nu(B) - \mu(B)$ by noting that $\nu(B) - \mu(B) = \mu(B^c) - \nu(B^c)$ which, after taking the supremum over B which are balls or complements of balls, obtain

$$\sup_B (\nu(B) - \mu(B)) = \sup_{B^c} (\mu(B^c) - \nu(B^c)) \leq \tilde{x} + \phi(\tilde{x})$$

as before. Since the supremum over balls and complements of balls will be larger than the supremum over balls, if $d_P(\mu, \nu) = x$, then $d_D(\mu, \nu) \leq \tilde{x} + \phi(\tilde{x})$ for all $\tilde{x} > x$. For right-continuous ϕ, the theorem follows by taking the limit as $\tilde{x}$ decreases to x. □

The Prokhorov and Wasserstein metrics. Huber (1981, p. 33) shows that

$$(d_P)^2 \leq d_W \leq 2\ d_P$$

for probability measures on a complete separable metric space whose metric d is bounded by 1. More generally, we show the following:

Theorem 2. *The Wasserstein and Prokhorov metrics satisfy*

$$(d_P)^2 \leq d_W \leq (\mathrm{diam}(\Omega) + 1)\ d_P$$

where $\mathrm{diam}(\Omega) = \sup\{d(x, y) : x, y \in \Omega\}$.

Proof. For any joint distribution J on random variables X, Y,

$$\begin{aligned} \mathbf{E}_J[d(X,Y)] &\leq \epsilon \cdot \Pr(d(X,Y) \leq \epsilon) + \mathrm{diam}(\Omega) \cdot \Pr(d(X,Y) > \epsilon) \\ &= \epsilon + (\mathrm{diam}(\Omega) - \epsilon) \cdot \Pr(d(X,Y) > \epsilon) \end{aligned}$$

If $d_P(\mu, \nu) \leq \epsilon$, we can choose a coupling so that $\Pr(d(X,Y) > \epsilon)$ is bounded by ϵ (Huber 1981, p. 27). Thus

$$\mathbf{E}_J[d(X,Y)] \leq \epsilon + (\mathrm{diam}(\Omega) - \epsilon)\epsilon \leq (\mathrm{diam}(\Omega) + 1)\epsilon.$$

Taking the infimum of both sides over all couplings, we obtain

$$d_W \leq (\mathrm{diam}(\Omega) + 1) d_P.$$

To bound Prokhorov by Wasserstein, use Markov's inequality and choose ϵ such that $d_W(\mu,\nu) = \epsilon^2$. Then

$$\Pr(d(X,Y) > \epsilon) \leq \frac{1}{\epsilon}\mathbf{E}_J[d(X,Y)] \leq \epsilon$$

where J is any joint distribution on X, Y. By Strassen's theorem (Huber 1981, Theorem 3.7), $\Pr(d(X,Y) > \epsilon) \leq \epsilon$ is equivalent to $\mu(B) \leq \nu(B^\epsilon) + \epsilon$ for all Borel sets B, giving $d_P^2 \leq d_W$. □

No such upper bound on d_W holds if Ω is not bounded. Dudley (1989, p. 330) cites the following example on $\mathbb{R}$. Let δ_x denote the delta measure at x. The measures $P_n := ((n-1)\delta_0 + \delta_n)/n$ converge to $P := \delta_0$ under the Prokhorov metric, but $d_W(P_n, P) = 1$ for all n. Thus Wasserstein metric metrizes weak convergence only on state spaces of bounded diameter.

The Wasserstein and Discrepancy metrics. The following bound can be recovered using the bounds through total variation (and is therefore not included on Figure 1), but we include this direct proof for completeness.

Theorem 3. *If Ω is finite,*

$$d_{\min} \cdot d_D \leq d_W$$

where $d_{\min} = \min_{x \neq y} d(x,y)$ over distinct pairs of points in Ω.

Proof. In the equivalent form of the Wasserstein metric, Equation (3), take

$$h(x) = \begin{cases} d_{\min} & \text{for } x \text{ in } B \\ 0 & \text{otherwise} \end{cases}$$

for B any closed ball. $h(x)$ satisfies the Lipschitz condition. Then

$$\begin{aligned} d_{\min} \cdot |\mu(B) - \nu(B)| &= \left| \int_\Omega h\, d\mu - \int_\Omega h\, d\nu \right| \\ &\leq d_W(\mu,\nu) \end{aligned}$$

and taking B to be the ball that maximizes $|\mu(B) - \nu(B)|$ gives the result. □

On continuous spaces, it is possible for d_W to converge to 0 while d_D remains at 1. For example, take delta measures δ_ϵ converging on δ_0.

The Total Variation and Discrepancy metrics. It is clear that

$$d_D \leq d_{TV} \tag{6}$$

since total variation is the supremum over a larger class of sets than discrepancy.

No expression of the reverse type can hold since the total variation distance between a discrete and continuous distribution is 1 while the discrepancy may be very small. Further examples are discussed in Section 5.

The Total Variation and Prokhorov metrics. Huber (1981, p. 34) proves the following bound for probabilities on metric spaces:

$$d_P \leq d_{TV}.$$

The Wasserstein and Total Variation metrics.

Theorem 4. *The Wasserstein metric and the total variation distance satisfy the following relation:*

$$d_W \leq \mathrm{diam}(\Omega) \cdot d_{TV}$$

where $diam(\Omega) = \sup\{d(x,y) : x, y \in \Omega\}$. *If* Ω *is a finite set, there is a bound the other way. If* $d_{\min} = \min_{x \neq y} d(x,y)$ *over distinct pairs of points in* Ω*, then*

$$d_{\min} \cdot d_{TV} \leq d_W. \tag{7}$$

Note that on an infinite set no such relation of the second type can occur because d_W may converge to 0 while d_{TV} remains fixed at 1. ($\min_{x \neq y} d(x,y)$ could be 0 on an infinite set.)

Proof. The first inequality follows from the coupling characterizations of Wasserstein and total variation by taking the infimum of the expected value over all possible joint distributions of both sides of:

$$d(X,Y) \leq \mathbf{1}_{X \neq Y} \cdot \mathrm{diam}(\Omega).$$

The reverse inequality follows similarly from:

$$d(X,Y) \geq \mathbf{1}_{X \neq Y} \cdot \min_{a \neq b} d(a,b).$$

□

The Hellinger and Total Variation metrics.

$$\frac{(d_H)^2}{2} \leq d_{TV} \leq d_H. \tag{8}$$

See LeCam (1969, p. 35).

The Separation distance and Total Variation. It is easy to show (see, e.g., Aldous and Diaconis (1987, p. 71)) that

$$d_{TV} \leq d_S. \tag{9}$$

As Aldous and Diaconis note, there is no general reverse inequality, since if μ is uniform on $\{1, 2, ..., n\}$ and ν is uniform on $\{1, 2, ..., n-1\}$ then $d_{TV}(\mu, \nu) = 1/n$ but $d_S(\mu, \nu) = 1$.

Relative Entropy and Total Variation. For countable state spaces Ω,

$$2\ (d_{TV})^2 \leq d_I.$$

This inequality is due to Kullback (1967). Some small refinements are possible where the left side of the inequality is replaced with a polynomial in d_{TV} with more terms; see Mathai and Rathie (1975, p. 110-112).

Relative Entropy and the Hellinger distance.

$$(d_H)^2 \leq d_I.$$

See Reiss (1989, p. 99).

The χ^2-distance and Hellinger distance.

$$d_H(\mu,\nu) \le \sqrt{2}(\ d_{\chi^2}(\mu,\nu))^{1/4}.$$

See Reiss (1989, p. 99), who also shows that if the measure μ is dominated by ν, then the above inequality can be strengthened:

$$d_H(\mu,\nu) \le (d_{\chi^2}(\mu,\nu))^{1/2}.$$

The χ^2-distance and Total Variation. For a countable state space Ω,

$$d_{TV}(\mu,\nu) = \frac{1}{2}\sum_{\omega\in\Omega}\frac{|\mu(\omega)-\nu(\omega)|}{\sqrt{\nu(\omega)}}\sqrt{\nu(\omega)} \le \frac{1}{2}\sqrt{d_{\chi^2}(\mu,\nu)}$$

where the inequality follows from Cauchy-Schwarz. On a continuous state space, if μ is dominated by ν the same relationship holds; see Reiss (1989, p. 99).

The χ^2-distance and Relative Entropy.

Theorem 5. *The relative entropy d_I and the χ^2-distance d_{χ^2} satisfy*

$$d_I(\mu,\nu) \le \log\left[1 + d_{\chi^2}(\mu,\nu)\right]. \tag{10}$$

In particular, $d_I(\mu,\nu) \le d_{\chi^2}(\mu,\nu)$.

Proof. Since log is a concave function, Jensen's inequality yields

$$d_I(\mu,\nu) \le \log\left(\int_\Omega (f/g)\ f\ d\lambda\right) \le \log\left(1 + d_{\chi^2}(\mu,\nu)\right) \le d_{\chi^2}(\mu,\nu),$$

where the second inequality is obtained by noting that

$$\int_\Omega \frac{(f-g)^2}{g}\ d\lambda = \int_\Omega \left(\frac{f^2}{g} - 2f + g\right)\ d\lambda = \int_\Omega \frac{f^2}{g}\ d\lambda - 1.$$

□

Diaconis and Saloff-Coste (1996, p. 710) derive the following alternate upper bound for the relative entropy in terms of both the χ^2 and total variation distances.

$$d_I(\mu,\nu) \le d_{TV}(\mu,\nu) + \frac{1}{2}d_{\chi^2}(\mu,\nu). \tag{11}$$

Weak convergence. In addition to using Figure 1 to recall specific bounds, our diagram there can also be used to discern relationships between topologies on the space of measures. For instance, we can see from the mutual arrows between the total variation and Hellinger metrics that they generate equivalent topologies. Other mutual arrows on the diagram indicate similar relationships, subject to the restrictions given on those bounds.

Moreover, since we know that the Prokhorov and Lévy metrics both metrize weak convergence, we can also tell which other metrics metrize weak convergence on which spaces, which we summarize in the following theorem:

Theorem 6. *For measures on $\mathbb{R}$, the Lévy metric metrizes weak convergence. Convergence under the discrepancy and Kolmogorov metrics imply weak convergence (via the Lévy metric). Furthermore, these metrics metrize weak convergence $\mu_n \to \nu$ if the limiting metric ν is absolutely continuous with respect to Lebesgue measure on $\mathbb{R}$.*

For measures on a measurable space Ω, the Prokhorov metric metrizes weak convergence. Convergence under the Wasserstein metric implies weak convergence.

Furthermore, if Ω is bounded, the Wasserstein metric metrizes weak convergence (via the Prokhorov metric), and convergence under any of the following metrics implies weak convergence: total variation, Hellinger, separation, relative entropy, and the χ^2-distance.

If Ω is both bounded and finite, the total variation and Hellinger metrics both metrize weak convergence.

This follows from chasing the diagram in Figure 1, noting the existence of mutual bounds of the Lévy and Prokhorov metrics with other metrics (using the results surveyed in this section) and reviewing conditions under which they apply.

4. Some Applications of Metrics and Metric Relationships

We describe some of the applications of these metrics in order to give the reader a feel for how they have been used, and describe how some authors have exploited metric relationships to obtain bounds for one metric via another.

The notion of *weak convergence* of measures is an important concept in both statistics and probability. For instance, when considering a statistic T that is a functional of an empirical distribution F, the "robustness" of the statistic under small deviations of F corresponds to the continuity of T with respect to the weak topology on the space of measures. See Huber (1981). The Lévy and Prokhorov metrics (and Wasserstein metric on a bounded state space) provide quantitative ways of metrizing this topology.

However, other distances that do not metrize this topology can still be useful for other reasons. The total variation distance is one of the most commonly used probability metrics, because it admits natural interpretations as well as useful bounding techniques. For instance, in (1), if A is any event, then total variation can be interpreted as an upper bound on the difference of probabilities that the event occurs under two measures. In Bayesian statistics, the error in an expected loss function due to the approximation of one measure by another is given (for bounded loss functions) by the total variation distance through its representation in equation (2).

In extending theorems on the ergodic behavior of Markov chains on discrete state spaces to general measurable spaces, the total variation norm is used in a number of results (Orey 1971, Nummelin 1984). More recently, total variation has found applications in bounding rates of convergence of random walks on groups (e.g., Diaconis (1988), Rosenthal (1995)) and Markov chain Monte Carlo algorithms (e.g., Tierney (1994), Gilks, Richardson and Spiegelhalter (1996)). Much of the success in obtaining rates of convergence in these settings is a result of the coupling characterization of the total variation distance, as well as Fourier bounds.

Gibbs (2000) considers a Markov chain Monte Carlo algorithm which converges in total variation distance, but for which coupling bounds are difficult to apply since the state space is continuous and one must wait for random variables to couple exactly. The Wasserstein metric has a coupling characterization that depends on the distance between two random variables, so one may instead consider only the time required for the random variables to couple to within ϵ, a fact exploited by Gibbs. For a related example with a discrete state space, Gibbs uses the bound (7) to obtain total variation bounds.

Like the total variation property (2), the Wasserstein metric also represents the error in the expected value of a certain class of functions due to the approximation of one measure by another, as in (3), which is of interest in applications in Bayesian statistics. The fact that the Wasserstein metric is a minimal distance of two random variables with fixed distributions has also led to its use in the study of distributions with fixed marginals (e.g., Rüschendorf, Schweizer and Taylor (1996)).

Because the separation distance has a characterization in terms of strong uniform times (like the coupling relationship for total variation), convergence of a Markov chain under the separation distance may be studied by constructing a strong uniform time for the chain and estimating the probability in the tail of its distribution. See Aldous and Diaconis (1987) for such examples; they also exploit inequality (9) to obtain upper bounds on the total variation.

Similarly, total variation lower bounds may be obtained via (6) and lower bounds on the discrepancy metric. A version of this metric is popular among number theorists to study uniform distribution of sequences (Kuipers and Niederreiter 1974); Diaconis (1988) suggested its use to study random walks on groups. Su (1998) uses the discrepancy metric to bound the convergence time of a random walk on the circle generated by a single irrational rotation. This walk converges weakly, but not in total variation distance because its n-th step probability distribution is finitely supported but its limiting measure is continuous (in fact, uniform). While the Prokhorov metric metrizes weak convergence, it is not easy to bound. On the other hand, for this walk, discrepancy convergence implies weak convergence when the limiting measure is uniform; and Fourier techniques for discrepancy allow the calculation of quantitative bounds. The discrepancy metric can be used similarly to study random walks on other homogeneous spaces, e.g., Su (2001).

Other metrics are useful because of their special properties. For instance, the Hellinger distance is convenient when working with convergence of product measures because it factors nicely in terms of the convergence of the components. Reiss (1989) uses this fact and the relationships (8) between the Hellinger and total variation distances to obtain total variation bounds. The Hellinger distance is also used in the theory of asymptotic efficiency (e.g., see LeCam (1986)) and minimum Hellinger distance estimation (e.g., see Lindsay (1994)). It is used throughout Ibragmiov and Has'minskii (1981) to quantify the rate of convergence of sequences of consistent estimators to their parameter. Kakutani (1948) gives a criterion (now known as the *Kakutani alternative*) using the Hellinger affinity to determine when infinite products of equivalent measures are equivalent; this has applications to stochastic processes and can be used to show the consistency of the likelihood-ratio test. See Jacod and Shiryaev (1987), Williams (1991) for applications.

Diaconis and Saloff-Coste (1996) use log-Sobolev techniques to bound the χ^2 convergence of Markov chains to their limiting distributions, noting that these also give total variation and entropy bounds. The χ^2-distance bears its name because in the discrete case it is the well-known χ^2 statistic used, for example, in the classic goodness-of-fit test, e.g., see Borovkov (1998, p. 184). Similarly, the Kolmogorov metric between a distribution function and its empirical estimate is used as the test statistic in the Kolmogorov-Smirnov goodness-of-fit test, e.g., see Lehmann (1994, p. 336).

Relative entropy is widely used because it is a quantity that arises naturally, especially in information theory (Cover and Thomas 1991). Statistical applications include proving central limit theorems (Linnik 1959, Barron 1986) and evaluating the loss when using a maximum likelihood versus a Bayes density estimate (Hartigan 1998). In the testing of an empirical distribution against an alternative, Borovkov (1998, p. 256) gives the relationship of the asymptotic behaviour of the Type II error to the relative entropy between the empirical and alternative distributions. In Bayesian statistics, Bernardo and Smith (1994, p. 75) suggest that relative entropy is the natural measure for the lack of fit of an approximation of a distribution when preferences are described by a logarithmic score function.

Up to a constant, the asymptotic behaviour of relative entropy, the Hellinger, and the χ^2-distance are identical when the ratio of the density functions is near 1 (Borovkov 1998, p. 178). These three distances are used extensively in parametric families of distributions to quantify the distance between measures from the same family indexed by different parameters. Borovkov (1998, pp. 180-181) shows how these distances are related to the Fisher information in the limit as the difference in the parameters goes to zero.

5. Rates of Convergence that Depend on the Metric

We now illustrate the ways in which the choice of metric can affect rates of convergence in one context: the convergence of a random walk to its limiting distribution. Such examples point to the need for practitioners to choose a metric carefully when measuring convergence, paying attention to that metric's qualitative and quantitative features. We give several examples of random walks whose qualitative convergence behavior depend strongly on the metric chosen, and suggest reasons for this phenomenon.

As a first basic fact, it is possible for convergence to occur in one metric but not another. An elementary example is the convergence of a standardized Binomial (n, p) random variable with distribution μ_n which converges to the standard normal distribution, ν, as $n \to \infty$. For all $n < \infty$, $d_{TV}(\mu_n, \nu) = 1$, while $d_D(\mu_n, \nu) \to 0$ as $n \to \infty$. In the random walk context, Su (1998) shows that a random walk on the circle generated by an irrational rotation converges in discrepancy, but not total variation. The latter fact follows because the n-th step probability distribution is finitely supported, and remains total variation distance 1 away from its continuous limiting distribution.

However, more interesting behavior can arise. Below, we cite a family of random walks on a product space, indexed by some parameter, which not only converges under each of total variation, relative entropy, and the χ^2-distance, but exhibits *different rates of convergence* as a function of the parameter.

We then cite another example of a family of walks that not only has different convergence rates under two different metrics, but also exhibits *qualitatively different* convergence behavior. This family exhibits a *cutoff phenomenon* under the first metric but only exponential decay under the second.

Example: convergence rates that depend on the metric. The following family of random walks show that convergence rates in total variation, relative entropy, and χ^2-distance may differ. Recall that a family of random walks, indexed by some parameter n, is said to *converge in $f(n)$ steps* using some metric/distance if that metric/distance can be uniformly bounded from above after $f(n)$ steps and is uniformly bounded away from zero before $f(n)$ steps.

Let $G = \mathbf{Z} \mod g$, a finite group with g elements. Then G^n is the set of all n-tuples of elements from G. Consider the following continuous-time random walk on G^n: start at $(0, 0, ..., 0)$ and according to a Poisson process running at rate 1, pick a coordinate uniformly at random and replace that coordinate by a uniformly chosen element from G. (Thus each coordinate is an independent Poisson process running at rate $\frac{1}{n}$.)

Su (1995) proves that if g grows with n exponentially according to $g = 2^n$, then the random walk on G^n described above converges in $2n \log n$ steps under the relative entropy distance, and $n^2 \log 2$ steps under the χ^2-distance. However, it converges in at most $n \log n$ steps under total variation.

In this example, the relative entropy and χ^2-distance are calculated by using properties of these metrics, while the total variation upper bound in this example is, in fact, calculated via its metric relationship (9) with the separation distance.

Why does the difference in rates of convergence occur? This is due to the fact that total variation distance, unlike relative entropy or the χ^2-distance, is insensitive to big or small elementwise differences when the total of those differences remains the same. For instance, consider the follow measures μ, ν on $\mathbf{Z}_{10}$. Let

$$\mu(i) = \begin{cases} 0.6 & \text{if } i = 0 \\ 0.1 & \text{if } i = 1,2,3,4 \\ 0 & \text{else} \end{cases} \qquad \nu(i) = \begin{cases} 0.2 & \text{if } i = 0,1,2,3,4 \\ 0 & \text{else} \end{cases}.$$

Let U be the uniform distribution on $\mathbf{Z}_{10}$. In this example $d_I(\mu, U) \approx 1.075$ and $d_I(\nu, U) = 0.693$, but the total variation distances are the same: $d_{TV}(\mu, U) = d_{TV}(\nu, U) = 0.5$. This is true because we could redistribute mass among the points on which μ exceeded ν without affecting the total variation distance.

Thus it is possible for an unbalanced measure (with large elementwise differences from uniform) to have the same total variation distance as a more balanced measure. The random walk on the product space has a natural feature which rewards total variation distance but hinders relative entropy— the randomization in each coordinate drops the variation distance quickly, but balancing the measure might take longer.

Example: qualitatively different convergence behavior. Chung, Diaconis and Graham (1987) study a family of walks that exhibits a *cutoff phenomenon* under total variation; the total variation distance stays near 1 but drops off sharply and steeply to zero after a certain cutoff point. The family of random walks is on the integers mod p where p is an odd number, with a randomness multiplier: the process is given by $X_0 = 0$ and $X_n = 2X_{n-1} + \epsilon_n \pmod p$ where the ϵ_i are i.i.d. taking values $0, \pm 1$ each with probability $1/3$. The stationary distribution for this process is uniform. Using Fourier analysis, Chung et al. (1987) show for the family of walks when $p = 2^t - 1$ that $O(\log_2 p \log\log_2 p)$ steps are sufficient and necessary for total variation convergence, for t a positive integer. There is, in fact, a cutoff phenomenon.

However, as proven in Su (1995, pp. 29-31), $O(\log p)$ steps are sufficient for convergence in discrepancy. Moreover, the convergence is qualitatively different from total variation, because there is no cutoff in discrepancy. Rather, there is only exponential decay.

Again, analysis under these two metrics sheds some light on what the walk is doing. For $p = 2^t - 1$, the "doubling" nature of the process keeps the walk supported very near powers of 2. The discrepancy metric, which accounts for the topology of the space, falls very quickly because the walk has spread itself out even though it is supported on a small set of values. However, the total variation does not "see" this spreading; it only falls when the support of the walk is large enough.

Thus, the difference in rates of convergence sheds some light on the nature of the random walk as well as the metrics themselves.

References

Aldous, D. and Diaconis, P. (1987). Strong uniform times and finite random walks, *Adv. in Appl. Math.* **8**: 69–97.

Barron, A. (1986). Entropy and central limit theorem, *Ann. Probab.* **14**: 336–42.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*, John Wiley & Sons, Chichester, England.

Borovkov, A. A. (1998). *Mathematical Statistics*, Gordon and Breach Science Publishers, Amsterdam. Translated from the Russian by A. Moullagaliev and revised by the author.

Chung, F., Diaconis, P. and Graham, R. L. (1987). Random walks arising in random number generation, *Ann. Probab.* **15**: 1148–1165.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*, John Wiley & Sons Inc., New York. A Wiley-Interscience Publication.

Csiszar, I. (1967). Information type measures of difference of probability distributions and indirect observations, *Studia Sci. Math. Hungar.* **2**: 299–318.

Diaconis, P. (1988). *Group Representations in Probability and Statistics*, Institute of Mathematical Statistics, Hayward, CA.

Diaconis, P. and Saloff-Coste, L. (1996). Logarithmic Sobolev inequalities for finite Markov chains, *Ann. Appl. Probab.* **6**(3): 695–750.

Diaconis, P. and Zabell, S. L. (1982). Updating subjective probability, *J. Amer. Statist. Assoc.* **77**: 822–830.

Dudley, R. M. (1968). Distances of probability measures and random variables, *Ann. Math. Statist.* **39**: 1563–1572.

Dudley, R. M. (1989). *Real Analysis and Probability*, Wadsworth & Brooks/Cole, Belmont, CA.

Gibbs, A. L. (2000). *Convergence of Markov Chain Monte Carlo Algorithms with Applications to Image Restoration*, PhD thesis, University of Toronto.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (eds) (1996). *Markov Chain Monte Carlo in Practice*, Chapman and Hall, London.

Hartigan, J. A. (1998). The maximum likelihood prior, *Ann. Statist.* **26**: 2083–103.

Hellinger, E. (1907). *Die Orthogonalinvarianten quadratischer Formen von unendlich vielen Variablen*, Dissertation, Göttingen.

Huber, P. J. (1981). *Robust Statistics*, John Wiley & Sons, New York.

Ibragmiov, I. A. and Has'minskii, R. Z. (1981). *Statistical Estimation: Asymptotic Theory*, Springer-Verlag, New York.

Jacod, J. and Shiryaev, A. N. (1987). *Limit Theorems for Stochastic Processes*, Springer-Verlag, Berlin.

Kakutani, S. (1948). On equivalence of infinite product measures, *Ann. of Math. (2)* **49**: 214–224.

Kolmogorov, A. N. (1933). Sulla determinazione empirica di una legge di distribuzione, *Giornale dell'Istituto Italiano degli Attuari* **4**: 83–91.

Kuipers, L. and Niederreiter, H. (1974). *Uniform Distribution of Sequences*, Wiley-Interscience [John Wiley & Sons], New York. Pure and Applied Mathematics.

Kullback, S. (1967). A lower bound for discrimination in terms of variation, *IEEE Trans. Inform. Theory* **4**: 126–127.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency, *Ann. Math. Statistics* **22**: 79–86.

Le Cam, L. and Yang, G. L. (1990). *Asymptotics in Statistics: Some Basic Concepts*, Springer-Verlag, New York.

LeCam, L. M. (1969). *Théorie Asymptotique de la Décision Statistique*, Les Presses de l'Université de Montréal, Montréal.

LeCam, L. M. (1986). *Asymptotic Methods in Statistical Decision Theory*, Springer-Verlag, New York.

Lehmann, E. L. (1994). *Testing Statistical Hypotheses*, 2nd edn, Chapman & Hall, New York.

Lévy, P. (1925). *Calcul des Probabilitiés*, Gauthiers-Villars, Paris.

Liese, F. and Vajda, I. (1987). *Convex Statistical Distances*, B.G. Teubner Verlagsgesellschaft, Leipzig.

Lindsay, B. G. (1994). Efficiency versus robustness: The case for minimum Hellinger distance and related methods, *Ann. Statist.* **22**: 1081–1114.

Lindvall, T. (1992). *Lectures on the Coupling Method*, John Wiley & Sons, New York.

Linnik, Y. V. (1959). An information-theoretic proof of the central limit theorem with the Lindeberg condition, *Theory Probab. Appl.* **4**: 288–99.

Lukacs, E. (1975). *Stochastic convergence*, second edn, Academic Press [Harcourt Brace Jovanovich Publishers], New York. Probability and Mathematical Statistics, Vol. 30.

Mathai, A. M. and Rathie, P. N. (1975). *Basic Concepts in Information Theory*, Wiley Eastern, New Delhi.

Nummelin, E. (1984). *General Irreducible Markov Chains and Nonnegative Operators*, Cambridge University Press, Cambridge.

Orey, S. (1971). *Lecture Notes on Limit Theorems for Markov Chain Transition Probabilities*, Van Nostrand Reinhold Co., London. Van Nostrand Reinhold Mathematical Studies, No. 34.

Petrov, V. V. (1995). *Limit theorems of probability theory*, The Clarendon Press Oxford University Press, New York. Sequences of independent random variables, Oxford Science Publications.

Prokhorov, Y. V. (1956). Convergence of random processes and limit theorems in probability theory, *Teor. Veroyatnost. i Primenen.* **1**: 177–238. English translation: *Theory Probab. Appl.* **1** (1956), 157-214.

Rachev, S. T. (1991). *Probability Metrics and the Stability of Stochastic Models*, John Wiley & Sons, Chichester, New York.

Reiss, R.-D. (1989). *Approximate Distributions of Order Statistics*, Springer-Verlag, New York.

Rosenthal, J. S. (1995). Convergence rates of Markov chains, *SIAM Rev.* **37**: 387–405.

Rüschendorf, L., Schweizer, B. and Taylor, M. D. (eds) (1996). *Distributions with Fixed Marginals and Related Topics*, Vol. 28 of *Lecture Notes – Monograph Series*, Institute of Mathematical Statistics, Hayward, California.

Shannon, C. E. (1948). A mathematical theory of communication, *Bell System Tech. J.* **27**: 379–423, 623–656.

Strassen, V. (1965). The existence of probability measures with given marginals, *Ann. Math. Statist* **36**: 423–439.

Su, F. E. (1995). *Methods for Quantifying Rates of Convergence for Random Walks on Groups*, PhD thesis, Harvard University.

Su, F. E. (1998). Convergence of random walks on the circle generated by an irrational rotation, *Trans. Amer. Math. Soc.* **350**: 3717–3741.

Su, F. E. (2001). Discrepancy convergence for the drunkard's walk on the sphere, *Electron. J. Probab.* **6**: no. 2, 20 pp. (electronic).

Szulga, A. (1982). On minimal metrics in the space of random variables, *Theory Probab. Appl.* **27**: 424–430.

Tierney, L. (1994). Markov chains for exploring posterior distributions, with discussion, *Ann. Statist.* **22**: 1701–1762.

Tierney, L. (1996). Introduction to general state-space Markov chain theory, *in* W. R. Gilks, S. Richardson and D. J. Spiegelhalter (eds), *Markov Chain Monte Carlo in Practice*, Chapman and Hall, London, pp. 59–74.

Williams, D. (1991). *Probability with Martingales*, Cambridge University Press, Cambridge.

Zolotarev, V. M. (1983). Probability metrics, *Theory Probab. Appl.* **28**: 278–302.

Department of Mathematics and Statistics, York University, 4700 Keele Street, Toronto, Ontario, Canada, M3J 1P3

E-mail address: `agibbs@mathstat.yorku.ca`

Department of Mathematics, Harvey Mudd College, Claremont, CA, U.S.A. 91711

E-mail address: `su@math.hmc.edu`