# A CLASS OF WASSERSTEIN METRICS FOR PROBABILITY DISTRIBUTIONS

Clark R. Givens and Rae Michael Shortt

**0. Introduction.** There are several natural metrics that one can place on spaces of probability distributions (or "laws"). These include total variation, Prohorov's $\rho$ metric, dual norms induced by spaces of Lipschitz functions, and the so-called Wasserstein distance. A discussion of these is to be found in Dudley [4], especially in Lectures 8, 18, and 20. See also [3].

The Wasserstein metric seems to have arisen first in connexion with the transport of mass problem. In a certain form, this dates back to 18th-Century work of Monge, but perhaps the first significant modern research was due to Kantorovich [8]. The realisation that the Wasserstein metric can be taken as a reasonable distance on spaces of random variables or probability distributions was first expressed in a paper of Kantorovich and Rubinstein [9], where the problem is put in the context of infinite-dimensional linear programming, and a duality theorem is proposed. This line of thought continues in Kemperman [10]. A general, abstract context for the metric is to be found in Szulga [16].

Although natural and far-reaching as a theoretical tool, the Wasserstein metric has a definite drawback: explicit calculation is difficult for most concrete examples. For distributions on the line, the problem is not severe, and there is a result of Vallander [17] to cover this case. In some unpublished work of Neveu and Dudley, the suggestion was made that a somewhat altered ($L^p$) version of the Wasserstein be considered. The present paper contains a calculation of the $L^2$ Wasserstein distance between arbitrary $n$-dimensional Gaussian distributions. The problem can be reduced to a Lagrange multiplier optimisation: this calculation forms §2 of the paper. Section 1 presents some general results concerning the family of $L^p$ Wassersteins for $1 \leqslant p \leqslant \infty$, whereas §3 concludes with a few open questions and speculations.

**1. The $L^p$ Wasserstein metrics.** Throughout this section, $(S, d)$ represents a complete, separable metric (Polish) space and 0 a fixed but arbitrarily chosen point in $S$. For each $p$ with $1 \leqslant p < \infty$, define $\mathfrak{M}_p = \mathfrak{M}_p(S)$ to be the collection of all probability measures (i.e. laws) $P$ on (the Borel sets of) $S$ for which

$$\int_S d^p(X, 0)\, dP(X)$$

is finite. Let $\mathfrak{M}_\infty(S)$ be the set of all laws on $S$ with bounded support. It is easy to show that the spaces $\mathfrak{M}_p$ do not depend on the choice of the point 0.

Let $P_1$ and $P_2$ be members of $\mathfrak{M}_p$ $(1 \leqslant p < \infty)$. The $L^p$ Wasserstein distance between $P_1$ and $P_2$ is defined by

Received January 9, 1984. Revision received March 16, 1984.
Michigan Math. J. 31 (1984).

(1) $$W_p(P_1,P_2)=\left(\inf\int d^p(X,Y)\,d\mu(X,Y)\right)^{1/p},$$

the infimum being taken over all $\mu$ in $D(P_1,P_2)$, the set of all laws $\mu$ on $S\times S$ with marginals $P_1$ and $P_2$. The case $p=1$ gives the usual Wasserstein. A simple triangle inequality shows that $W_p$ is finite for laws in $\mathfrak{M}_p$. The $L^\infty$ distance is defined as

(2) $$W_\infty(P_1,P_2)=\inf\|d(X,Y)\|_\infty^{(\mu)},$$

where the superscription indicates that the usual $L^\infty$ norm is taken with respect to $\mu$. Again, the infimum is over all $\mu$ in $D(P_1,P_2)$ and is clearly finite for laws in $\mathfrak{M}_\infty$.

The lemma and proposition that follow will prove quite handy in our analysis of $W_p$.

LEMMA 1. *Let $\mu_0,\mu_1,\mu_2,\ldots$ be a sequence of laws on $S\times S$, each of whose marginals is a member of $\mathfrak{M}_p(S)$, $1\leqslant p<\infty$. If $\mu_n\to\mu_0$ (weakly) as $n\to\infty$, then*

(3) $$\liminf_{n\to\infty}\int d^p(X,Y)\,d\mu_n(X,Y)\geqslant\int d^p(X,Y)\,d\mu_0(X,Y).$$

*Proof.* We write

(4) $$\int d^p(X,Y)\,d\mu_n=\int_0^\infty \mu_n\{(X,Y):d^p(X,Y)>r\}\,dr.$$

An application of Fatou's lemma and the usual Portmanteau theorem for weak convergence yields the result. □

PROPOSITION 1. *Given laws $P_1$ and $P_2$ in $\mathfrak{M}_p$ $(1\leqslant p\leqslant\infty)$, the infimum in* (1) *is attained for some law $\mu$ in $D(P_1,P_2)$.*

*Demonstration.* We take first the case where $p<\infty$. Let $\mu_1,\mu_2,\ldots$ be a sequence of laws in $D(P_1,P_2)$ such that

(5) $$\int d^p(X,Y)\,d\mu_n<W_p^p(P_1,P_2)+1/n.$$

Noting that $D(P_1,P_2)$ is compact for weak convergence, we may produce a subsequence $\mu_{n(k)}$ converging as $k\to\infty$ to a law $\mu$ in $D(P_1,P_2)$. Using Lemma 1 and (5), one sees that

$$\left(\int d^p(X,Y)\,d\mu\right)^{1/p}\leqslant W_p(P_1,P_2).$$

The infimum is thus attained at $\mu$.

For the case $p=\infty$, again choose $\mu_1,\mu_2,\ldots$ in $D(P_1,P_2)$ with

$$\|d(X,Y)\|_\infty^{(\mu_n)}<W_\infty(P_1,P_2)+1/n.$$

Put $B_n=\{(X,Y)\in S\times S:d(X,Y)\leqslant W_\infty(P_1,P_2)+1/n\}$. Then $\mu_n(B_m)=1$ for all $n\geqslant m$. As before, let $\mu_{n(k)}\to\mu$ as $k\to\infty$. Since each $B_m$ is closed, $\mu(B_m)=1$, and so $\mu(B_1 \cap B_2 \cap \cdots) = 1$, proving that

$$\|d(X,Y)\|_{\infty}^{(\mu)} \leqslant W_{\infty}(P_1,P_2).$$

Hence they are equal. □

PROPOSITION 2. *The Wasserstein functions $W_p$ are metrics on the sets $\mathfrak{M}_p$ for $1 \leqslant p \leqslant \infty$.*

*Demonstration.* The only point requiring a certain subtlety is the verification of the triangle inequality. We check the case $p<\infty$; the situation for $p=\infty$ is similar. For notational ease, put $S_1=S_2=S_3=S$. Given $\epsilon>0$ and $P_1$, $P_2$, $P_3$ in $\mathfrak{M}_p$, let $\mu_{12}$ and $\mu_{23}$ be laws on $S_1\times S_2$ and $S_2\times S_3$ with marginals $P_1$, $P_2$, $P_3$ for which

$$W_p(P_1,P_2) = \left(\int d^p(X,Y)\, d\mu_{12}\right)^{1/p} \text{ and}$$

$$W_p(P_2,P_3) = \left(\int d^p(Y,Z)\, d\mu_{23}\right)^{1/p};$$

we have used Proposition 1. Then let $\mu$ be a law on $S_1\times S_2\times S_3$ with bivariate marginals $\mu_{12}$ and $\mu_{23}$. For a discussion of the existence of such a law (mathematical folklore), see Theorem 5 in Shortt [13]. Let $\mu_{13}$ be the marginal of $\mu$ on $S_1\times S_3$. Then, applying Minkowski's Inequality in $L^p(\mu)$,

$$\begin{aligned} W_p(P_1,P_3) &\leqslant \left(\int d^p(X,Z)\, d\mu_{13}\right)^{1/p} \\ &\leqslant \left(\int (d(X,Y)+d(Y,Z))^p\, d\mu(X,Y,Z)\right)^{1/p} \\ &\leqslant \left(\int d^p(X,Y)\, d\mu_{12}\right)^{1/p} + \left(\int d^p(Y,Z)\, d\mu_{23}\right)^{1/p} \\ &= W_p(P_1,P_2) + W_p(P_2,P_3). \end{aligned}$$ □

As was mentioned in the introduction, a fair number of metrics have been placed on spaces of probability measures. We recall two such presently. For any real function $f$ on $S$, define

$$\|f\|_{\infty} = \sup_{x\in S}|f(x)| \qquad \text{(supremum norm)}$$

$$\|f\|_L = \sup_{x\neq y} \frac{|f(x)-f(y)|}{d(x,y)} \qquad \text{(Lipschitz semi-norm)}$$

$$\|f\|_{BL} = \|f\|_{\infty} + \|f\|_L.$$

If $\nu$ is a finite signed measure on $S$, then we define its dual Lipschitz norm as $\|\nu\|_L^* = \sup\{|\int f\,d\nu| : \|f\|_L \leqslant 1\}$.

THEOREM (Kantorovich–Rubinstein). *For any laws $P_1$ and $P_2$ in $\mathfrak{M}_1(S)$, one has $W_1(P_1,P_2) = \|P_1 - P_2\|_L^*$.*

*Proof.* See Fernique [5]. For related results, extensions, and partial proofs, consult [4], [9], and [10].

For any two laws $P_1$ and $P_2$ on $S$, the $\beta$ distance is defined by

$$\beta(P_1,P_2)=\sup\left\{\int f\,d(P_1-P_2):\ \|f\|_{BL}\leqslant 1\right\}.$$

See Dudley ([4, Lecture 8] or [2, Theorem 9]) for a proof that $\beta$ is a complete metric inducing the topology of weak convergence.

PROPOSITION 3. *For any two laws $P_1$ and $P_2$ on $S$, the following inequalities obtain:*

(1) $\beta(P_1,P_2)\leqslant W_1(P_1,P_2)$; *and*

(2) $W_p(P_1,P_2)\leqslant W_{p'}(P_1,P_2)$ *for* $1\leqslant p\leqslant p'\leqslant\infty$.

*In addition,*

(3) $\lim_{p\to\infty} W_p(P_1,P_2)=W_\infty(P_1,P_2)$.

*Demonstration.* (1) follows immediately from the Kantorovich–Rubinstein theorem and the definition of $\beta$.

For (2), note that the case $p'=\infty$ is trivial, whereas Jensen's Inequality applied to $f(X,Y)=d^p(X,Y)$ and the convex function $\phi(x)=x^{p'/p}$ yields the result for $p'<\infty$.

For (3), put $L=\lim_{p\to\infty} W_p(P_1,P_2)$. From (2), $L\leqslant W_\infty(P_1,P_2)$. To prove equality, we exhibit the case $W_\infty(P_1,P_2)<\infty$; the other case invites a similar argument. Given $\epsilon>0$, let $\mu_n$ be a law in $D(P_1,P_2)$ with

$$\left(\int d^n(X,Y)\,d\mu_n(X,Y)\right)^{1/n}=W_n(P_1,P_2).$$

As in previous arguments, let $\mu_{n(k)}\to\mu$ weakly as $k\to\infty$.

*Case* 1: $M=\|d(X,Y)\|_\infty^{(\mu)}$ is finite. Then put $U=\{(X,Y): d(X,Y)>M-\epsilon\}$ and note that $\mu(U)>0$. Since $U$ is open, one has $\mu_{n(k)}(U)>\mu(U)$ for all large $k$. Then

$$W_{n(k)}(P_1,P_2)\geqslant\mu_{n(k)}(U)^{1/n(k)}(M-\epsilon)\geqslant\mu(U)^{1/n(k)}(M-\epsilon).$$

Letting $k\to\infty$ gives $L\geqslant M-\epsilon$. Let $\epsilon$ evaporate: $L=M$ as desired.

*Case* 2: $\|d(X,Y)\|_\infty^{(\mu)}$ is infinite. Then replace $M$ in Case 1 with an arbitrary positive integer and set $U=\{(X,Y): d(X,Y)>M\}$. The same reasoning applies. □

Proposition 3 implies that on their common domain, the topologies induced by the metrics $W_p$ are ordered in strength. For example, convergence of laws $P_n\to P$ for any $W_p$ implies the usual weak convergence. In general, the $W_p$ give rise to distinct topologies, as the following shows.

EXAMPLE. Let $S=\mathbf{R}$, the real line under its usual metric. For each $x\in\mathbf{R}$, let $\delta_x$ be the point mass at $x$. If $\xi$ is a real-valued random variable with law $\mathcal{L}(\xi)=P$, then $W_p(P,\delta_0)=\|\xi\|_p$, the usual $L^p$ norm of $\xi$. By choosing $\xi_n\to 0$ in mean of order $p$ but not order $p'>p$, one sees that the topologies induced by the $W_p$ are indeed distinct.

The analogy between the spaces $(\mathfrak{M}_p, W_p)$ and the Banach spaces $L^p$ is extensive. An explicit link is made in the case where $S=\mathbf{R}$. Then the map sending an $L^p$ element $\xi$ to its law $\mathcal{L}(\xi)$ is contractive from $L^p$ to $\mathfrak{M}_p$, that is, $W_p(\mathcal{L}(\xi), \mathcal{L}(\eta)) \leqslant \|\xi - \eta\|_p$. Using this idea, one proves the following.

PROPOSITION 4. *If $(S,d)$ is bounded, then the spaces $\mathfrak{M}_p$ coincide for $1 \leqslant p \leqslant \infty$ and the metrics $W_p$ induce the topology of weak convergence for $1 \leqslant p < \infty$.*

*Demonstration.* Suppose that $P_n \to P$ as $n \to \infty$ for the usual weak topology. Then by the Skorohod Embedding Theorem (Skorohod [14] or Dudley [4]), there is a probability space $(\Omega, Q)$ and $S$-valued random variables $\xi$ and $\xi_n$, $n=1,2,\ldots$, on $\Omega$ with $\xi_n \to \xi$ $Q-$a.s. and $\mathcal{L}(\xi)=P$, $\mathcal{L}(\xi_n)=P_n$. Then $d(\xi_n, \xi)$, $n=1,2,\ldots$, are uniformly bounded real functions converging to 0 $Q-$a.s. It follows from Vitali's Convergence Theorem (Hewitt and Stromberg [7, 13.38]) that $d(\xi_n,\xi) \to 0$ in $L^p(\Omega, Q)$ for $1 \leqslant p < \infty$. Since $W_p(P_n, P) \leqslant \|d(\xi_n, \xi)\|_p$, the proposition follows. □

However, note that the topology induced by $W_\infty$ will, in general, be rather stronger than that of weak convergence. There is a convenient description of $W_\infty$ which bears comparison with another oft-used metric. Given laws $P_1$ and $P_2$ on $S$, define Prohorov's metric $\rho(P_1, P_2)$ by

$$\rho(P_1, P_2) = \inf\{\epsilon > 0 : P_1(A) \leqslant P_2(A^\epsilon) + \epsilon \text{ all } A\},$$

where $A^\epsilon$ represents the $\epsilon$-neighbourhood of the Borel set $A$, that is, $A^\epsilon = \{x \in S : d(x,a) \leqslant \epsilon \text{ for some } a \in A\}$. As is well known, $\rho$ metrises convergence of laws: see Dudley [4, Lecture 8]. By comparison, one has the following.

PROPOSITION 5. *If $P_1$ and $P_2$ are laws in $\mathfrak{M}_\infty(S)$, then*

$$W_\infty(P_1, P_2) = \inf\{\epsilon > 0 : P_1(A) \leqslant P_2(A^\epsilon) \text{ all } A\}.$$

*Demonstration.* Note that $W_\infty(P_1, P_2) \leqslant \epsilon$ if and only if there is some $\mu$ in $D(P_1, P_2)$ with $\mu(B_\epsilon) = 1$, where $B_\epsilon = \{(X,Y) : d(X,Y) \leqslant \epsilon\}$ is a closed subset of $S \times S$. As a consequence of Strassen [15, Theorem 11] or Shortt [13, Theorem 1], such a $\mu$ exists if and only if $(A \times S) \cap B_\epsilon \subset (S \times B) \cap B_\epsilon$ implies $P_1(A) \leqslant P_2(B)$ for all Borel subsets $A, B$ of $S$. But whenever $(A \times S) \cap B_\epsilon \subset (S \times B) \cap B_\epsilon$, then also $A^\epsilon \subset B$. Thus $W_\infty(P_1, P_2) \leqslant \epsilon$ if and only if $P_1(A) \leqslant P_2(A^\epsilon)$ for all Borel sets $A$. The proposition follows. □

As an aside, we note that although $W_\infty$ induces a strong topology, it is not in general comparable with the topology of convergence in total variation.

We conclude this section with a result parallel to the classical Riesz–Fischer Theorem.

PROPOSITION 6. *For each $p$ $(1 \leqslant p \leqslant \infty)$, the metric spaces $(\mathfrak{M}_p, W_p)$ are complete.*

*Demonstration.* We take first the case $p < \infty$. Let $P_1, P_2, \ldots$ be a sequence of laws in $\mathfrak{M}_p$, Cauchy for $W_p$. Then from Proposition 3, the sequence $P_1, P_2, \ldots$

is also Cauchy for the $\beta$ metric and so converges weakly to some law $P$ on $S$. Let $\delta_0$ be the point mass at 0. Then the inequality

$$|W_p(P_n,\delta_0)-W_p(P_m,\delta_0)|\leqslant W_p(P_n,P_m)$$

implies that the sequence

$$W_p(P_n,\delta_0)=\left(\int d^p(X,0)\,dP_n(X)\right)^{1/p}$$

is Cauchy and therefore bounded. Proceeding as in Lemma 1, we find that

$$W_p(P,\delta_0)\leqslant \liminf W_p(P_n,\delta_0)$$

is finite, so that $P$ is in $\mathfrak{M}_p(S)$.

*Claim:* $W_p(P_n,P)\to 0$ as $n\to\infty$. Using Proposition 1, we select laws $\mu_{nm}$ in $D(P_n,P_m)$ for which

$$W_p(P_n,P_m)=\left(\int d^p(X,Y)\,d\mu_{nm}\right)^{1/p}.$$

Given $\epsilon>0$, choose $N$ large so that for all $m,n\geqslant N$, $W_p(P_n,P_m)\leqslant\epsilon$. For each $n\geqslant N$, consider the sequence $\mu_{nm}$ for $m=n,n+1,n+2,\ldots$. It is uniformly tight and so contains a subsequence $\mu_{nm(k)}$ converging weakly to a law $\mu_n$ with marginals $P_n$ and $P$. Applying Lemma 1 to this subsequence, we find that

$$\epsilon\geqslant\liminf_{k\to\infty} W_p(P_n,P_{m(k)})\geqslant W_p(P_n,P)$$

for all $n\geqslant N$. The claim is established.

The case $p=\infty$ proceeds analogously, and the proof is both straightforward and omitted. □

**2. The $L^2$ Wasserstein for Gaussian measures.** Let $P_1$ and $P_2$ be Gaussian measures on $\mathbf{R}^n$ with means $\vec{m}_1$ and $\vec{m}_2$ and non-singular covariance matrices $M_1$ and $M_2$ respectively. This notation will remain fixed for the rest of the present section, which is devoted to a proof of the the following.

PROPOSITION 7. *The $L^2$ Wasserstein distance $W_2(P_1,P_2)$ is given by*

$$\sqrt{\|\vec{m}_1-\vec{m}_2\|^2+\operatorname{tr}(M_1)+\operatorname{tr}(M_2)-2\operatorname{tr}[(\sqrt{M_1}M_2\sqrt{M_1})^{1/2}]}. \tag{6}$$

To begin the calculation, we first reduce to the case where $\vec{m}_1=\vec{m}_2=0$. Let $X$ and $Y$ be $\mathbf{R}^n$-valued random variables with $\mathcal{L}(X)=P_1$ and $\mathcal{L}(Y)=P_2$. Then $W_2(P_1,P_2)$ is the infimum of $\sqrt{E\|X-Y\|^2}$ taken over all possible joint distributions of $X$ and $Y$. Put $\xi=X-\vec{m}_1$ and $\eta=Y-\vec{m}_2$; also set $Q_1=\mathcal{L}(\xi)$ and $Q_2=\mathcal{L}(\eta)$. Simply note that $E\|X-Y\|^2=E\|\xi-\eta\|^2+\|\vec{m}_1-\vec{m}_2\|^2$; it follows that

$$W_2(P_1,P_2)=\sqrt{\|\vec{m}_1-\vec{m}_2\|^2+W_2(Q_1,Q_2)}.$$

Thus we can and do assume that $\vec{m}_1=\vec{m}_2=0$. Note that this reduction does not require $P_1$ and $P_2$ to be Gaussian.

The next step is to show that in calculating the infimum in (1), we may restrict ourselves entirely to Gaussian measures.

LEMMA 2. *The infimum in* (1) *is attained for a Gaussian law* $\nu$ *in* $D(P_1, P_2)$.

*Proof.* It follows from Proposition 1 that the infimum in (1) is attained for some law $\mu$ in $D(P_1, P_2)$. Simply let $\nu$ be a Gaussian measure on $\mathbf{R}^{2n}$ with the same covariance matrix as $\mu$. Then also $\nu \in D(P_1, P_2)$, and

$$\int d^2(X,Y)\,d\nu = \int d^2(X,Y)\,d\mu = W_2^2(P_1,P_2). \tag{7}$$ □

Thus, in the search for optimal $\mu$ in (1), it suffices to consider only mean 0 Gaussians on $\mathbf{R}^n \times \mathbf{R}^n$, or what is the same, their corresponding covariance matrices $A$. The condition that $\mu$ have marginals $P_1$ and $P_2$ is equivalent to the requirement that $A$ have the block form

$$A = \begin{bmatrix} M_1 & K \\ K^T & M_2 \end{bmatrix}, \tag{8}$$

where $K$ is some $n \times n$ matrix. Then

$$\int d^2(X,Y)\,d\mu(X,Y) = \operatorname{tr}(M_1) + \operatorname{tr}(M_2) - 2\operatorname{tr}(K). \tag{9}$$

The problem thus reduces to the finding of a matrix $K$ that minimises (9) subject to the constraint that (8) be non-negative definite. For each $m$, let $D(m)$ and $D_0(m)$ denote the classes of positive definite and non-negative definite $m \times m$ matrices, respectively. The covariance matrix $A$ from (8) admits a factorisation

$$A = \begin{bmatrix} M_1^{1/2} & 0 \\ K^T M_1^{-1/2} & I \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & S \end{bmatrix} \begin{bmatrix} M_1^{1/2} & M_1^{-1/2}K \\ 0 & I \end{bmatrix},$$

where $S$ is the Schur complement $S = M_2 - K^T M_1^{-1} K$. Note that since $M_1, M_2 \in D(n)$, the square root $M_1^{1/2}$ and its inverse are well-defined members of $D(n)$. From the factorisation, one sees that $A \in D_0(2n)$ if and only if $S \in D_0(n)$. This condition on $S$ defines the set $\mathcal{P}$ of possible $K$ over which the infimum is to be taken in (9).

Define $\phi$ on $\mathcal{P}$ by the rule $\phi(K) = M_2 - K^T M_1^{-1} K$ and let

$$f(K) = \operatorname{tr}(M_1) + \operatorname{tr}(M_2) - 2\operatorname{tr}(K).$$

By comparing the minima of $f$ restricted to the fibres $\phi^{-1}(S)$, $S \in \phi(\mathcal{P})$, we shall show that the infimum in (9) occurs for some $K$ in $\phi^{-1}(0)$.

For $S \in \phi(\mathcal{P})$, the fibre over $S$ is the set of $K$ for which $K^T M_1^{-1} K = M_2 - S$. Since $M_1^{-1} \in D(n)$, we note that $M_2 - S \in D_0(n)$. Thus, if $\operatorname{rank}(M_2 - S) = r$, we have the spectral decomposition

$$M_2 - S = U\Lambda^2 U^T = U_r \Lambda_r^2 U_r^T, \tag{10}$$

where $\Lambda^2 = \operatorname{diag}(\lambda_1^2, \ldots, \lambda_r^2, 0, \ldots, 0) = \Lambda_r^2 \oplus 0$ and $U = [U_r, U_{n-r}]$ is a matrix of corresponding orthonormal eigenvectors.

Let $\Lambda_r$ denote any of the invertible matrices $\operatorname{diag}(\pm\lambda_1, \ldots, \pm\lambda_r)$, held fixed for the moment. From (10), we conclude that

$$(M_1^{-1/2} K U_r \Lambda_r^{-1})^T (M_1^{-1/2} K U_r \Lambda_r^{-1}) = I \quad (r \times r \text{ identity}),$$

and so

$$KU_r = \sqrt{M_1}\,\mathcal{O}_r\Lambda_r$$

with $\mathcal{O}_r$ an arbitrary $n\times r$ matrix satisfying $\mathcal{O}_r^T\mathcal{O}_r = I$ (an "$r$-frame"). Moreover, $KU_{n-r}=0$, since $M_1^{-1}\in D(n)$. Thus

$$K = KUU^T = KU_rU_r^T = \sqrt{M_1}\,\mathcal{O}_r\Lambda_rU_r^T,$$

and the fibre $\phi^{-1}(S)$ is parametrized by the $r$-frames $\mathcal{O}_r$. On this fibre, the function $f$ now assumes the form

$$f(\mathcal{O}_r) = \operatorname{tr}(M_1) + \operatorname{tr}(M_2) - 2\operatorname{tr}(\mathcal{O}_r^T\sqrt{M_1}\,U_r\Lambda_r). \tag{11}$$

Consider now the problem of minimizing a function $f(\mathcal{O}) = \text{constant} - 2\operatorname{tr}(\mathcal{O}^TB)$ subject to the constraint $\mathcal{O}^T\mathcal{O} = I$, where $\mathcal{O}, B$ are $n\times r$, and $B$ is of full rank. If $\mathcal{O} = [v_1,\dots,v_r]$, then in a Lagrange multiplier approach to the minimisation, the equivalent set of constraints $v_i^Tv_j = \delta_{ij}$, $i\leqslant j$, would be incorporated into an auxiliary function as

$$\sum_{i,j} c_{ij}(v_i^Tv_j - \delta_{ij}), \quad c_{ij} = c_{ji}.$$

Thus, the appropriate form at the matrix level is $\operatorname{tr}[C(\mathcal{O}^T\mathcal{O} - I)]$, with $C$ a symmetric $r\times r$ Lagrange multiplier matrix. The auxiliary function is now defined by

$$F(\mathcal{O}, C) = \text{constant} - 2\operatorname{tr}(\mathcal{O}^TB) + \operatorname{tr}[C(\mathcal{O}^T\mathcal{O} - I)].$$

At the critical points of $F$, defined by $F_{\mathcal{O}} = F_C = 0$, we have $\mathcal{O}C = B$ and $\mathcal{O}^T\mathcal{O} = I$. Since $\mathcal{O}$ and $B$ are of rank $r$, $C^{-1}$ exists, and $\mathcal{O} = BC^{-1}$. Thus, $C$ is some square root of $B^TB$. At the critical points, $f = \text{constant} - 2\operatorname{tr}(\sqrt{B^TB})$, and it is clear that $f$ is minimised by taking the positive definite square root of $B^TB$. We note also that $\operatorname{tr}(\sqrt{B^TB}) = \operatorname{tr}(\sqrt{BB^T})$.

If we now apply the results of this Lagrange multiplier analysis to the case at hand in (11), where $B = \sqrt{M_1}\,U_r\Lambda_r$, then

$$\begin{aligned}(f|_{\phi^{-1}(S)})_{\min} &= \operatorname{tr}(M_1) + \operatorname{tr}(M_2) - 2\operatorname{tr}\sqrt{\sqrt{M_1}\,U_r\Lambda_r^2U_r^T\sqrt{M_1}} \\ &= \operatorname{tr}(M_1) + \operatorname{tr}(M_2) - 2\operatorname{tr}\sqrt{\sqrt{M_1}(M_2 - S)\sqrt{M_1}},\end{aligned}$$

using (10) for the second equality.

And now a comparison of the minima of $f$ over the various fibres is accomplished by appeal to the following min-max result.

THEOREM (Courant-Fischer). *Let $M$ be a symmetric $n\times n$ matrix with eigenvalues $\mu_1\geqslant\cdots\geqslant\mu_n$ and Rayleigh quotient*

$$R(\xi) = \frac{\xi^TM\xi}{\xi^T\xi} = R(\xi/|\xi|).$$

*Given an arbitrary $k$-dimensional subspace $V_k$ in $\mathbf{R}^n$, let $S_{n-k-1}$ be the unit sphere in the orthocomplement $V_k^\perp$. Then*

$$\mu_{k+1} = \min_{V_k}\max_{\xi\in S_{n-k-1}} R(\xi).$$

*Proof.* See Lancaster [11, 3.6.1, p. 116].

For a given Schur complement $S$, let $\mu_1^2(S) \geqslant \cdots \geqslant \mu_n^2(S)$ denote the eigenvalues of $\sqrt{M_1}(M_2 - S)\sqrt{M_1}$. Since $S \in D_0(n)$, we have

$$\xi^T \sqrt{M_1}(M_2 - S)\sqrt{M_1}\,\xi \leqslant \xi^T \sqrt{M_1} M_2 \sqrt{M_1}\,\xi.$$

Because this inequality persists when $\xi$ is constrained to various subspaces of $\mathbf{R}^n$, we conclude from the Courant–Fischer Theorem that

$$\mu_j^2(S) \leqslant \mu_j^2(0) \quad j = 1, \ldots, n$$

and consequently that

$$f_{\min} = (f|_{\phi^{-1}(0)})_{\min} = \operatorname{tr}(M_1) + \operatorname{tr}(M_2) - 2\operatorname{tr}[(\sqrt{M_1} M_2 \sqrt{M_1})^{1/2}], \tag{12}$$

which establishes Proposition 7. □

We conclude with a few remarks concerning Proposition 7. As expected of a metric, the right-hand side of (12) is symmetric in $M_1$ and $M_2$ (by the earlier observation that $\operatorname{tr}(\sqrt{B^T B}) = \operatorname{tr}(\sqrt{BB^T})$) and vanishes when $M_1 = M_2$. Also, since the metric property of $W_2$ has been independently established in Proposition 2, it is possible to obtain a matrix inequality on three positive definite matrices $M_1$, $M_2$, $M_3$ by appealing to the triangle inequality for $W_2$. Note also that $f_{\min}$ reduces to $f_{\min}^0 = \operatorname{tr}[(\sqrt{M_1} - \sqrt{M_2})^2]$ when $M_1$ and $M_2$ commute. Lastly, we compute the special cases $n = 1$ and $n = 2$.

COROLLARY. *Let $P_1$ and $P_2$ be mean* 0 *Gaussian measures on* $\mathbf{R}^n$ *with covariance matrices $M_1$ and $M_2$. Proposition* 7 *implies that* (1) *for $n = 2$,*

$$W_2(P_1, P_2) = \sqrt{\operatorname{tr}(M_1) + \operatorname{tr}(M_2) - 2[\operatorname{tr}(M_1 M_2) + 2\sqrt{\det(M_1 M_2)}]^{1/2}};$$

*and* (2) *for $n = 1$,*

$$W_2(P_1, P_2) = |\sqrt{M_1} - \sqrt{M_2}|.$$

*Proof.* (2) is clear, whereas (1) follows from the formula $(\operatorname{tr}(\sqrt{B}))^2 = \operatorname{tr}(B) + 2\sqrt{\det B}$ for $B$ in $D(2)$: this is obtained by taking traces in the characteristic polynomial for $\sqrt{B}$. □

To conclude, we note that (6) in Proposition 7 is also valid in the case where $M_1$ and $M_2$ are singular. Similar arguments apply, but require some tedious checking of cases.

**3. Questions and conjectures.** A deeper analysis of the set of optimal $\mu$ occurring in (1) is probably warranted. Are the optimal $\mu$ always singular with respect to the product measure $P_1 \otimes P_2$ when $P_1$ and $P_2$ are continuous? Several such questions could be asked. A matter of great interest to probabilists is the a.s. convergence of empirical measures $P_n$ to their underlying law $P$. Results of Fortet and Mourier [6] combined with the Kantorovich–Rubinstein Theorem ensure that for $P$ in $\mathcal{M}_1(S)$, the associated empirical measures $P_n$ converge a.s. to $P$ for the metric $W_1$. Convergence in $W_p$ for $1 < p < \infty$ may also be proved, and will be treated in a later paper.

Finally, we pose the problem of explicit calculation of the $L^2$ Wasserstein in the case of Gaussian measures on infinite-dimensional linear spaces, in particular

Hilbert space. A corresponding infinite-dimensional Lagrange multiplier analysis would no doubt yield the same formulae.

The authors would like to thank the referee for some helpful comments. He has also pointed out an earlier attempt (cf. D. C. Dowson and B. V. Landau, *The Frechet distance between multivariate normal distributions,* J. Multivariate Analysis 12 (1982), 450–455) to calculate the $L^2$ Wasserstein distance between two multivariate Gaussian probabilities. He notes that their result is valid only when the corresponding covariance matrices commute.

## REFERENCES

1. H. Bauer, *Probability theory and elements of measure theory,* Holt, Rinehart and Winston, New York, 1972.
2. R. M. Dudley, *Convergence of Baire measures,* Studia Math. 27 (1966), 251–268.
3. ———, *Distances of probability measures and random variables,* Ann. Math. Statist. 39 (1968), 1563–1572.
4. ———, *Probabilities and metrics,* Lecture Notes Series, No. 45, Aarhus Universitet, Aarhus, 1976.
5. X. Fernique, *Sur le théorème de Kantorovitch–Rubinstein dans les espaces polonaise.* Seminar on probability, XV (Strasbourg, 1979/1980), 6–10, Lecture Notes in Math., 850, Springer, Berlin, 1981.
6. R. Fortet and E. Mourier, *Convergence de la répartition empirique vers la répartition théorique,* Ann. Sci. École Norm. Sup. (3) 70 (1953), 267–285.
7. E. Hewitt and K. Stromberg, *Real and abstract analysis,* Springer, New York, 1965.
8. L. Kantorovich, *On the translocation of masses,* C. R. Acad. Sci. URSS (N.S) 37 (1942), 199–201.
9. L. Kantorovich and G. Rubinstein, *On a space of completely additive functions* (Russian), Vestnik Leningrad. Univ. 13 (1958), 52–59.
10. J. H. B. Kemperman, *On the role of duality in the theory of moments, Semi-infinite programming and applications,* Lecture Notes in Economics and Mathematical Systems, 215 (1981), 63–92.
11. P. Lancaster, *Theory of matrices,* Academic Press, New York, 1969.
12. R. M. Shortt, *Strassen's marginal problem in two or more dimensions,* Z. Wahrsch. Verw. Gebiete 64 (1983), 313–325.
13. ———, *Universally measurable spaces: an invariance theorem and diverse characterizations,* Fund. Math. 121 (1983), 35–42.
14. A. V. Skorohod, *Limit theorems for stochastic processes,* Theory Probab. Appl. 1 (1956), 261–290.
15. V. Strassen, *The existence of probability measures with given marginals,* Ann. Math. Statist. 36 (1965), 423–439.
16. A. Szulga, *On minimal metrics in the space of random variables.* Theory Probab. Appl. 27 (1982), 424–430.
17. S. S. Vallander, *Calculation of the Wasserstein distance between probability distributions on the line,* Theory Probab. Appl. 18 (1973), 784–786.

Department of Mathematical & Computer Sciences
Michigan Technological University
Houghton, Michigan 49931