Permanents, Transportation Polytopes and Positive Definite Kernels on Histograms

Marco Cuturi
Institute of Statistical Mathematics, Tokyo, Japan
cuturi@ism.ac.jp

Abstract

For two integral histograms $r = (r_1, \ldots, r_d)$ and $c = (c_1, \ldots, c_d)$ of equal sum N, the Monge-Kantorovich distance $d_{\mathrm{MK}}(r, c)$ between r and c parameterized by a $d \times d$ distance matrix T is the minimum of all costs $< F, T >$ taken over matrices F of the transportation polytope $U(r, c)$. Recent results suggest that this distance is not negative definite, and hence, through Schoenberg's well-known result, $\exp(-\frac{1}{t} d_{\mathrm{MK}})$ may not be a positive definite kernel for all $t > 0$. Rather than using directly d_{MK} to define a similarity between r and c, we propose in this paper to investigate kernels on r and c based on the whole transportation polytope $U(r, c)$. We prove that when r and c have binary counts, which is equivalent to stating that r and c represent clouds of points of equal size, the permanent of an adequate Gram matrix induced by the distance matrix T is a positive definite kernel under favorable conditions on T. We also show that the volume of the polytope $U(r, c)$, that is the number of integral transportation plans, is a positive definite quantity in r and c through the Robinson-Schensted-Knuth correspondence between transportation matrices and Young Tableaux. We follow by proposing a family of positive definite kernels related to the generating function of the polytope through recent results obtained separately by A. Barvinok on the one hand, and C.Berg and A.J. Duran on the other hand. We finally present preliminary results led on a subset of the MNIST database to compare clouds of points through the permanent kernel.

1 Introduction

Defining meaningful kernels on histograms and clouds of points – and more generally positive measures on an arbitrary space $\mathcal{X}$ – is an important topic in the field of kernel methods, as it is directly related to the definition of kernels for structured objects seen as bags-of-components. Since the latter representations are frequently used by practitioners in applications, notably images seen as histograms of colors, texts as bags-of-words or sequences as groups of subsequences, research has been active in this field recently.

In the early applications of kernel methods to complex data structures, histograms were often treated as vectors, and used as such with the standard Gaussian or polynomial kernels [Joachims, 2002]. More adequate positive definite kernels which exploit their specificity have been proposed since. Namely, kernels which take into account the fact that histograms are vectors with nonnegative coordinates [Hein and Bousquet, 2005], and whose sum may be normalized to one, that is cast as discrete probability measures and treated under the light of information geometry [Lafferty and Lebanon, 2005; Lebanon, 2006]. Since such histograms are usually defined on bins which are not equally dissimilar, as is for instance the case with color or amino-acid histograms, further kernels which may take into account an a priori inter-bin similarity where subsequently proposed [Kondor and Jebara, 2003; Cuturi *et al.*, 2005; Hein and Bousquet, 2005].

In this context, a well-known distance for probability measures on a space $\mathcal{X}$ which takes explicitly into account the geometry of $\mathcal{X}$ is the optimal transportation distance [Villani, 2001], which is usually known as the Monge-Kantorovich (MK) or Wasserstein distance. This distance is also popular in the computer vision community [Rubner *et al.*, 2000] under the name of the earth movers' distance. However, preliminary findings [Naor and Schechtman, 2005] suggest that the MK distance is not negative definite, and cannot thus be used directly to define positive definite kernels, through the Schoenberg theorem[1] and negative exponentiation for instance. Although some approximations of the distance have been used so far to define positive definite kernels in vision applications [Grauman and Darrell, 2004], we propose in this paper to consider not only the optimal transport plan, but the whole of the transportation polytope to characterize the similarity of two histograms r and c.

This idea is rooted in the approach of [Vert *et al.*, 2004] to define a positive definite kernel for strings derived from a set of string manipulations which may map a string m_1 to another string m_2, namely sequences of deletions, substitutions and insertions of tokens, known as alignments. Vert

[1] The Schoenberg theorem[Berg *et al.*, 1984, Theorem 3.2.2], states that if a function ψ is a negative definite kernel, equivalently that $-\psi$ is conditionally positive definite [Schölkopf and Smola, 2002], then for all $t > 0$, $\exp(-t\psi)$ is a positive definite kernel. The Gaussian kernel is for instance based on the fact that for two vectors x, y, $\|x - y\|^2$ is negative definite.

et al. [2004] consider all possible alignments π between m_1 and m_2 and associate to each of these alignments π a score $S(\pi)$ which quantifies how efficiently the sequence π aligns successive tokens of m_1 and m_2. Rather than considering the score $S(\pi^\star)$ of the optimal alignment $\pi^\star$, known as the Smith-Waterman score in the context of biological sequences, Vert *et al.* [2004] propose to define a positive definite kernel between m_1 and m_2 through the sum $\sum_\pi e^{\beta S(\pi)}$, $\beta > 0$, which in their experimental setting provides a much better performance. Intuitively, the latter sum can be interpreted as the generating function of the set of all alignments π, which may give it more discriminative power than the simple use of the extremum $S(\pi^\star)$.

In the context of this paper, the set of all alignments π is played by the set of all $d \times d$ transportation matrices $F \in U(r, c)$ between two discrete histograms $r = (r_1, \ldots, r_d)$ and $c = (c_1, \ldots, c_d)$ of equal sum N; the analog of the BLOSUM distance matrices between amino-acids used in Vert *et al.* [2004] is an arbitrary $d \times d$ negative definite distance matrix T between the bins of the histograms, and finally, the cost $\pi(\sigma)$ becomes simply the Frobenius dot-product $< F, T >$. As is also the case in [Vert *et al.*, 2004], the family of convolution kernels introduced by Haussler [1999] plays an important role in our proofs, notably for the permanent kernel for clouds of points introduced in Section 2. We then propose in Section 3 a kernel between histograms by only taking into account the volume of $U(r, c)$, and we show that this result is a natural consequence of the Robinson-Schensted-Knuth correspondence between transportation matrices and generalized Young tableaux. Finally, inspired by a recent construction obtained by Barvinok [2005] and through a lemma by Berg and Duran [2004], we show in Section 4 that a weighted version of the generating function of $U(r, c)$ can be used to define positive definite kernels. We close the paper with Section 5 by discussing implementation and computational issues brought forward by these kernels, as well as preliminary experimental results led on a subset of the MNIST database of handwritten digits.

2 Permanent kernel for clouds of points

We define in this section a kernel for two clouds of points $x = \{x_1, \ldots, x_n\}$ and $y = \{y_1, \ldots, y_n\}$ in a space $\mathcal{X}$ endowed with a kernel κ, through a kernel on arbitrary sequence representations $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$, which is by definition invariant under reordering of these terms. Recall that for a $n \times n$ matrix $M = [m_{ij}]$, the permanent of M, per M is defined as the quantity

$$\operatorname{per} M = \sum_{\sigma \in S_n} \prod_{i=1}^{n} m_{i\sigma(i)}$$

where σ spans the symmetric group S_n, that is the set of all permutations of $\{1, \ldots, n\}$. Note that the definition of the permanent of a matrix differs from that of its determinant in that the signatures of the permutations are not taken into account. The permanent of a matrix is also invariant under any permutation of columns or rows of this matrix, a fact that we use in the proof of Proposition 1 below.

Proposition 1 *Let $\mathcal{X}$ be a set endowed with a kernel κ and $\mathcal{X}_n$ the set of clouds of points of $\mathcal{X}$ of cardinal n, that is $\{X = \{x_1, \ldots, x_n\}, x_i \in \mathcal{X}\}$. Let $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_n\} \in \mathcal{X}_n$. Then*

$$k_{\text{per}} : (X, Y) \mapsto \operatorname{per}([\kappa(x_i, y_j)]_{1 \le i, j \le n}) \qquad (1)$$

is a positive definite kernel on $\mathcal{X}_n \times \mathcal{X}_n$.

Proof. We first prove the result for two sequences $\mathbf{x} = (x_1, \ldots, x_n), \mathbf{y} = (y_1, \ldots, y_n)$. Given a permutation σ we write $\mathbf{x}_\sigma$ for the sequence $(x_{\sigma(1)}, \ldots, x_{\sigma(n)})$. The proof now follows from the work of Haussler [1999] on convolution kernels. Namely, we consider the equivalence relation $\mathcal{R}$ between two sequences $\mathbf{x}, \mathbf{y}$ where $\mathbf{x}\mathcal{R}\mathbf{y}$ if and only if there exists a permutation $\sigma \in S_n$ such that $\mathbf{x}_\sigma = \mathbf{y}$. Consider now the kernel k for two sequences

$$k((x_1, \ldots, x_n), (y_1, \ldots, y_n)) = \prod_{1 \le i \le n} \kappa(x_i, y_i).$$

which we use to define the convolution kernel K, which is positive definite by definition,

$$\begin{aligned} K(\mathbf{x}, \mathbf{y}) &= \sum_{\mathbf{u} \in \mathcal{R}^{-1}\mathbf{x}} \sum_{\mathbf{v} \in \mathcal{R}^{-1}\mathbf{y}} k(\mathbf{u}, \mathbf{v}) \\ &= \sum_{\sigma_u \in S_n} \sum_{\sigma_v \in S_n} k(\mathbf{x}_{\sigma_u}, \mathbf{y}_{\sigma_v}) \\ &= \sum_{\sigma_u \in S_n} \sum_{\sigma_v \in S_n} \prod_{1 \le i \le n} \kappa(x_{\sigma_u(i)}, y_{\sigma_v(i)}) \\ &= \sum_{\sigma_u \in S_n} \operatorname{per}[\kappa(x_i, y_j)] = n! \operatorname{per}[\kappa(x_i, y_j)], \end{aligned}$$

hence the positive definiteness of k_{per} used on clouds of points X, Y represented through any arbitrary pair of sequences $\mathbf{x}, \mathbf{y}$.■

Suppose for interpretation purposes that the kernel k can be written as $k(x, y) = e^{-d(x,y)}$ where d is an Hilbertian metric [Hein and Bousquet, 2005] on $\mathcal{X}$, that is there exists a mapping ϕ from $\mathcal{X}$ to an arbitrary Hilbert space $\mathcal{H}$ such that $d(x, y) = \|\phi(x) - \phi(y)\|$. In that case the permanent can be interpreted as the sum over all possible matchings σ of the weight of each total transport scheme e^{-d_σ} where $d_\sigma = \sum_i d(x_i, y_{\sigma(i)})$. The quantity d_σ stands for the total transport cost between the two clouds-of-points given the transport plan σ is selected, taken in a feature space $\mathcal{H}$, as illustrated in Figure 1.

Note finally that a possible way to define kernels for two clouds X and Y of sizes n and m respectively is to consider the sum of the pairwise kernels of all their respective subsets of size $d \le \min(n, m)$, that is

$$k_{\text{per}}(X, Y) = \sum_{x_1^d \in \mathbf{x}} \sum_{y_1^d \in \mathbf{y}} k_{\text{per}}(x_1^d, y_1^d).$$

3 The volume of the transport polytope as a kernel for marginals

We write $\mathbb{N} = \{0, 1, \ldots\}$ for the set of nonnegative integers, and consider now integral histograms (or margins as in the

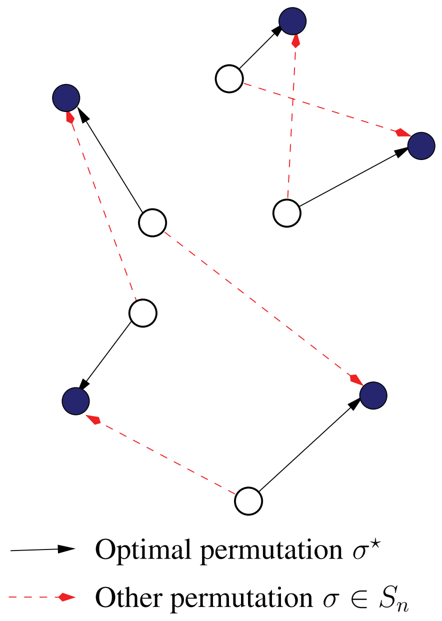
Optimal permutation $\sigma^\star$
Other permutation $\sigma \in S_n$

Figure 1: By considering the permanent of the matrix $e^{-t_{ij}}$ where $t_{ij} = d(x_i, y_j)$, the pairwise distances between white and dark points, the permanent kernel explicitly considers the costs of all possible matchings between the points of x and y, and not only the optimal permutation $\sigma^\star$.

corresponding statistical literature [Diaconis and Gangolli, 1995]) with identical overall sum and dimension, that is elements of the simplex lattice

$$\Sigma_{d,N} \overset{\text{def}}{=} \{r = (r_i) \in \mathbb{N}^d, \sum_{i=1}^{d} r_i = N\}.$$

We consider the polytope of transport matrices between r and c, restricted to integral matrices, that is

$$U(r,c) = \{F \in \mathbb{N}_{d\times d} \,|\, F\mathbf{1}_d = r,\ F^\top \mathbf{1}_d = c\},$$

where $\mathbf{1}_d$ is the d-dimensional vector of ones. We recall that the optimal transportation cost from r to c is a symmetric function in r and c which is defined as the result of the optimization

$$d_{\text{MK}}(r,c) \overset{\text{def}}{=} \min_{F\in U(r,c)} < F, T >,$$

where $T \in \mathbb{R}_{d,d}$ is an arbitrary distance matrix between bins, and for two square matrices U and V we use the Frobenius dot-product $< U, V >= \operatorname{tr}(UV^\top)$. Note that the optimal plan

$$F^\star \overset{\text{def}}{=} \operatorname{argmin}_{F\in U(r,c)} < F, T >$$

can be computed through standard linear-programming methods in polynomial time in d and it is known that $F^\star$ is a vertex of the polytope. We will reconsider the cost parameter T in the next section, and focus for the rest of this section on the volume $|U(r,c)|$ of the polytope $U(r,c)$, that is the total number of integral transportation plans. We introduce first the concept of *semi-standard Young tableaux*.

For two partitions u, v of an integer n, a semi-standard Young tableau of shape u and weight v is a diagram of shape u containing v_1 ones, v_2 twos, etc., arranged to be weakly right increasing in rows and strictly increasing down columns. Consider for instance for $n = 15$, the semi-standard tableau

$$\begin{array}{llllll} 1 & 1 & 1 & 2 & 4 & 7 \\ 2 & 3 & 3 & 5 & & \\ 3 & 4 & 6 & 6 & & \\ 6 & & & & & \end{array}$$

of shape $u = (6, 4, 4, 1)$ and weights $v = (3, 2, 3, 2, 1, 3, 1)$. Given two partitions u and v, the Kostka number $K_{u,v}$ is equal to the number of semi-standard Young tableaux of shape u and weight v.

Proposition 2 *The kernel k_{vol} on $\Sigma_{d,N}$ defined as*

$$k_{\text{vol}}(r,c) = |U(r,c)|$$

is symmetric positive definite.

Proof. The proof can be derived either from the theory of symmetric functions or from the Robinson-Schensted-Knuth (RSK) correspondence [Knuth, 1970], with both approaches mentionned in [Diaconis and Gangolli, 1995]. We recall briefly the second proof. The RSK bijective correspondence states that to every matrix $M \in U(r,c)$ corresponds one and only pair of semi-standard Young tableaux of identical shape and weights r and c respectively. We hence have that, summing over all possible partitions η of N used as shapes for the Young tableaux,

$$|U(r,c)| = \sum_{\eta} K_{\eta,\,r} K_{\eta,\,c}\ ,$$

which is sufficient to prove that the volume $|U(r,c)|$ satisfies Mercer's condition.■

4 Weighted generating functions

The volume of the transportation polytope is a special case of the evaluation of the generating function f of $U(r,c)$ [Barvinok, 2006] on a given cost matrix T, in that case the null matrix, where f is more generally defined as

$$f(T) = \sum_{F\in U(r,c)} e^{-<F,T>}.$$

The computation of f for general polytopes, notably the transportation one, is a subject of extensive research, with significative developments carried out in recent years and summarized in [Loera *et al.*, 2004]. The generating function can also be expressed as the total weight of $U(r,c)$ if we use the terminology of [Barvinok, 2005] by setting $w_{ij} = e^{-t_{ij}}$. Note that for binary histograms, that is marginals which may either take 1 or 0 values, $U(r,c)$ is known as the Birkhoff polytope and $f(T)$ corresponds in this case to the permanent kernel defined in Proposition 1 with κ set to e^{-T}. In the general case where r and c may not have binary counts, the volume $|U(r,c)|$ can thus be regarded as a similarity between r and c based exclusively on combinatorial properties, regardless of any prior knowledge T on the distance between the d

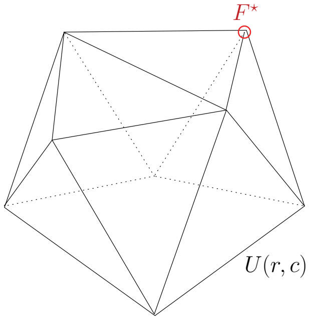

Figure 2: Rather than only consider the minimal value for $< F, T >$ reached on the vertex $F^\star$, we propose to use the same cost criterion evaluated on the whole polytope $U(r, c)$.

bins. On the other hand, the optimal plan $F^\star$ corresponding to the Monge-Kantorovich distance $d_{\mathrm{MK}} =< F^\star, T >$ takes into account such an information but does not reflect the information carried out by the distribution of the costs found in the whole polytope. Hence, having in mind Figure 2 we propose to define valid kernels k_φ which consider both criterions, that is we consider the distribution of the cost $< F, T >$ over the whole polytope $U(r, c)$. We do so by introducing weighted versions of the generating function through a function $\varphi : U(r, c) \to \mathbb{R}$, defining

$$k_\varphi : (r, c) \mapsto \sum_{F \in U(r,c)} \varphi(F) e^{-<T,F>}.$$

Note that the generating functionis recovered when $\varphi = 1$. Although the generating function might be a good candidate for a kernel between r and c, we do not know at this moment whether its evaluations on arbitrary matrices T are positive definite functions of r and c. We provide instead a family of functions φ which ensures this condition:

Proposition 3 *Given that T is such that $[e^{-t_{ij}}]$ is positive semidefinite, and having defined for $0 \leq a < 2$ the weight function $\varphi_a : \mathbb{N}_{d\times d} \to \mathbb{R}$ as*

$$\varphi_a(F) = \prod_i \frac{(2f_{ii})!^a}{f_{ii}!} \prod_{i \neq j} f_{ij}!^{2a-1},$$

we have that k_{φ_a} is a positive definite kernel on $\Sigma_{d,N}$.

The symmetry of k_{φ_a} is ensured by the symmetry of T and φ_a since $U(c, r) = U(r, c)^\top$ and we have that

$$\varphi_a(F) e^{-<F,T>} = \varphi_a(F^\top) e^{-<F^\top,T>}.$$

We prove the positive-definiteness of k_{φ_a} using the following 3 lemmas, which are motivated by a recent characterization carried out by Barvinok [2005] of the generating function of the transport polytope in terms of random permanents.

Lemma 4 *Let T be a $d \times d$ cost matrix such that $[e^{-t_{ij}}]_{1\leq i,j\leq d}$ is positive semidefinite and $\gamma = (\gamma_1, \ldots, \gamma_d)$ a sequence of nonnegative real numbers. For $r, c \in \Sigma_{d,N}$ define the $N \times N$ block matrix A as*

$$A = [A_{i,j}]_{1\leq i,j\leq d}$$

where each block $A_{i,j}$ is the $r_i \times c_j$ rectangular matrix with all coefficients set to the constant $\gamma_i \gamma_j e^{-t_{i,j}}$. Then

$$k_{\gamma,T} : (r, c) \mapsto \frac{\operatorname{per} A}{r_1! \cdots r_d! c_1! \cdots c_n!}$$

is a positive definite kernel on $\Sigma_{d,N} \times \Sigma_{d,N}$.

Proof. we first map each marginal r and c to the corresponding sequences

$$\tilde{r} = (\underbrace{1, \ldots, 1}_{r_1 \text{ times}}, \underbrace{2, \ldots, 2}_{r_2 \text{ times}}, \ldots, \underbrace{d, \ldots, d}_{r_d \text{ times}})$$

and $\tilde{c}$, and define the positive definite kernel

$$\kappa(i, j) = \gamma_i \gamma_j e^{-t_{i,j}},$$

for the kernel indexed on $\{1, \ldots, d\} \times \{1, \ldots, d\}$. We then have using Proposition 1 that:

$$k_{\gamma,T}(r, c) = k_{\mathrm{per}}(\tilde{r}, \tilde{c}) \cdot \frac{1}{r_1! \cdots r_d! c_1! \cdots c_d!}.$$

Since $(r, c) \mapsto \frac{1}{r_1! \cdots r_d!} \times \frac{1}{c_1! \cdots c_d!}$ is trivially positive definite, so is $k_{\gamma,T}$ as the product of two positive definite kernels.■

We use the following lemma to turn the randomized setting proposed in [Barvinok, 2005] into a sum of positive definite kernels:

Lemma 5 (Berg, Duran) *For each $0 < \alpha \leq 2$, the sequence $(n!)^\alpha, n \in \mathbb{N}$ is a determinate Stieltjes moment sequence, that is there exists a unique nonnegative measure μ_α on $[0, \infty[$ such that $\int_0^\infty x^n \mu_\alpha(x) = (n!)^\alpha$ for $n \in \mathbb{N}$.*

We refer to [Berg and Duran, 2004] for a proof of this result, and more generally to the reference [Berg *et al.*, 1984] for the exposition of the moment problems and their relationship with harmonic analysis on semigroups. Note that in the case where $\alpha = 1$, μ_1 is the standard exponential density, and for $\alpha = 0$ the measure μ_0 can be simply defined as the dirac mass on 1.

Lemma 6 (Barvinok) *Let $0 \leq a < 2$ and suppose $\gamma = (\gamma_1, \ldots, \gamma_d)$ is distributed as a sequence of independent random variables with identical law μ_a. Through the identity*

$$k_{\varphi_a} = E[k_{\gamma,T}(r, c)]$$

we have that k_{φ_a} is positive definite.

Proof. $E[k_{\gamma,T}(r, c)]$ is trivially positive definite as a sum of positive definite kernels. We follow Barvinok's proof to prove the equality, with a slight modification: Barvinok considers standard exponential variable γ_{ij} arranged in a $d \times d$ matrix, while we consider here a sequence of independent random variables $\gamma = (\gamma_1, \ldots, \gamma_d)$ which all follow law μ_a.

Let us consider matrix A defined in Lemma 4. For every permutation σ of S_N let

$$h_\sigma = \prod_{k=1}^N a_{k\sigma(k)}$$

be the corresponding term in $\mathrm{per}\, A$. Hence

$$E[\mathrm{per}\, A] = \sum_{\sigma \in S_N} E[h_\sigma].$$

Following Barvinok, with every permutation σ we associate a transport plan $D = D(\sigma)$ of $U(r, c)$ called the *pattern* of σ, as follows. Namely $D = [d_{ij}]_{1 \leq i,j \leq d}$ where

$$d_{ij} = \sum_{k=1}^{N} \mathbf{1}(\tilde{r}_k = i)\mathbf{1}(\tilde{c}_{\sigma(k)} = j),$$

that is d_{ij} is the number of indices $k \in \{1, \ldots, N\}$ such that $(k, \sigma(k))$ is in the (i, j) block of A. Note that $D : \sigma \mapsto D(\sigma) \in U(r, c)$ is surjective, but not bijective as we see below. For h_σ, we thus have, through Lemma 5 that

$$\begin{aligned} E[h_\sigma] &= E[\prod_{i,j}(e^{-t_{ij}}\gamma_i\gamma_j)^{d_{i,j}}] = \prod_{i,j} e^{-t_{ij}d_{ij}} E[\gamma_i^{d_{i,j}}\gamma_j^{d_{i,j}}] \\ &= \prod_{i,j} e^{-t_{ij}d_{ij}} \prod_i (2d_{ii})!^a \prod_{i \neq j} (d_{ij}!)^{2a}. \end{aligned}$$

At this point we follow exactly Barvinok's proof. Barvinok proves that the number of permutations σ of S_N which admit D as a pattern is

$$C_{r,c,D} = \frac{r_1! \cdots r_d! c_1! \cdots c_d!}{\prod_{i,j} d_{ij}!},$$

yielding

$$E[\mathrm{per}\, A] = \sum_{D \in U(r,c)} C_{r,c,D} \prod_{i,j} e^{-t_{ij}d_{ij}} \prod_i (2d_{ii})!^a \prod_{i \neq j} d_{ij}!^{2a},$$

and hence

$$\begin{aligned} E[k_{\gamma,T}(r, c)] &= \sum_{D \in U(r,c)} \varphi_a(D) \prod_{i,j} e^{-t_{ij}d_{ij}} \\ &= \sum_{D \in U(r,c)} \varphi_a(D) e^{-<T,D>} = k_{\varphi_a}(r, c), \end{aligned}$$

which concludes the proof.■

Note that when $a = 0$ we obtain the Fisher-Yates distribution on transportation matrices [Diaconis and Gangolli, 1995], that is

$$\varphi_0(F) = \prod_{i,j} \frac{1}{f_{ij}!},$$

whereas the case $a = \frac{1}{2}$ yields a weight which only depends on the diagonal elements of the transport plan,

$$\varphi_{\frac{1}{2}}(F) = \prod_i \frac{\sqrt{(2f_{ii})!}}{f_{ii}!}.$$

5 Discussion and Experiments

We discuss in this section complexity issues which may arise when trying to compute the kernels presented above, and we present preliminary results on a pattern recognition task which involves comparing clouds of points.

5.1 Computational issues

To handle clouds of points through Proposition 1 requires the computation of the permanent of a $n \times n$ matrix, which is a notoriously difficult problem in combinatorics. Millions of computations of such kernel evaluations, which are usually required to fill in Gram matrices, may not be tractable at the moment when the number n of points exceeds twenty to thirty points. However, and in the case where the kernel κ is bounded between 0 and 1, recent advances[2] in the computation of approximations of the permanent through Sequential Monte Carlo (SMC) techniques [Jerrum *et al.*, 2004] yield a complexity of the order of $n^7 \log^4 n$. This is still problematic for large n, but we believe that for clouds of points of small size the permanent might be a useful kernel, with the ability of quantifying complex relationships through the power of combinatorics. We propose below in our experiments to compare 2000 images of handwritten digits by sampling artificially 20 black pixels among each image, and compare these clouds of points through the permanent of the pairwise kernels for the points in each cloud. Another issue in that case arises from the numerical stability of the computation of the permanent when the values for κ might be too small, and we do not have an adequate answer to this problem other than simple cross-validation to obtain reasonable entries.

In the more general case of histograms, both the computation of the volume $|U(r, c)|$ and the integration of k_{φ_0}, which corresponds to the Fisher-Yates distribution, as well as that of $k_{\varphi_{1/2}}$, may be computed through SMC sampling methodologies presented in recent works [Chen *et al.*, 2006]. For the volume only, exact calculations through toolboxes such as `LattE` [Loera *et al.*, 2004] are possible, but only tractable for very low dimensions. Diaconis and Gangolli [1995] propose ad-hoc numerical approximations when d is small and N is large,

$$|U(r, c)| = \frac{\Gamma(dk)(N + \frac{1}{2}d^2)^{(d-1)^2}}{\Gamma(d)^d k^d} \prod_{i=1}^{d} (\bar{r}_i)^{d-1} (\bar{c}_i)^{k-1}$$

where

$$\begin{aligned} w &= \frac{1}{1 + d^2/2N}, \quad k = \frac{d+1}{d \sum \bar{r}_i^2} - \frac{1}{d}, \\ \bar{r}_i &= \frac{1-w}{d} + \frac{wr_i}{N}, \quad \bar{c}_j = \frac{1-w}{d} + \frac{wc_j}{N}. \end{aligned}$$

Although these expressions might be symmetrized by averaging $\frac{1}{2}(|U(r, c)| + |U(c, r)|)$, their positive definiteness may not be guaranteed and has yet to be tested on datasets.

5.2 Experiments

Following the previous work of Kondor and Jebara [2003], we have conducted experiments on the first 2000 images (28×28 pixels) of the MNIST database of handwritten digits, with approximately 200 images for each digit. For each image X^i, we randomly sample a set $\{x_1^i, \ldots, x_{20}^i\}$ of 20 distinct black points in the image, that is pixels with an intensity superior to 190 represented as points of the square

[2] which we have not used in our experiments.

$[0,1] \times [0,1]$, and perform a multiclass classification to classify any new image as one of the 10 digits. We do so by applying a simple one-vs-all strategy on 10 classifiers, namely support vector machines, using the Spider toolbox. This setting makes it particularly difficult for most common kernels to compare the images and we consider here two different approaches: first the permanent kernel of Equation (1), second a Gaussian kernel taken between the two images seen as 28×28 dimensional vectors, preliminarily smoothed through a smoothing-kernel κ on the pixels, which yields actually a simple summation over Gram matrices as described for instance in [Borgwardt *et al.*, 2006]. The kernel κ used to compute both the permanent as in Equation (1), and to smooth the image in the second case was set to be the Gaussian kernel between pixels $\kappa(x,y) = \exp(-\|x-y\|^2/\sigma^2)$ with a width σ spanning values $0.1, 0.2$ and 0.3. The considered kernels are thus

$$\begin{aligned} k_{\mathrm{per}}(X^i, X^j) &= \mathrm{per}[\kappa(x_r^i, x_s^j)]_{r,s}, \\ k_{\mathrm{gaussian}}(X^i, X^j) &= \exp(-\sum_{r,s} \kappa(x_r^i, x_s^j)), \end{aligned}$$

and we use their normalized counterpart, that is using $\tilde{k}(x,y) = k(x,y)/\sqrt{k(x,x)k(y,y)}$ instead of k in our experiments. We report the cross-validation errors for these settings for σ taken over 5-fold cross validations in Table 1, which show that the permanent kernels appear as a robust although costly alternative to the smoothed kernel in this preliminary experiment.

σ	Gaussian	Permanent
.1	34.3 (± 1.4)	32.3 (± 1.2)
.2	33.45 (± 1.0)	31.3 (± 1.3)
.3	37.3 (± 1.0)	33.2 (± 1.2)

Table 1: Misclassification rate expressed in percents for the 2 considered kernels along with their standard errors averaged over cross-validation folds.

References

Alexander Barvinok. Enumerating contingency tables via random permanents, 2005. arXiv.org:math/0511596.

Alexander Barvinok. The complexity of generating functions for integer points in polyhedra and beyond. In *Proceedings of the International Congress of Mathematicians, Madrid*, 2006. to appear.

Christian Berg and Antonio J. Duran. A transformation from hausdorff to Stieltjes moment sequences. *Arkiv för matematik*, 42:239–257, 2004.

Christian Berg, Jens Peter Reus Christensen, and Paul Ressel. *Harmonic Analysis on Semigroups*. Number 100 in Graduate Texts in Mathematics. Springer Verlag, 1984.

Karsten M. Borgwardt, Arthur Gretton, Malte J. Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alexander J. Smola. Integrating structured biological data by kernel maximum mean discrepancy. In *ISMB (Supplement of Bioinformatics)*, volume 22, pages 49–57, 2006.

Yuguo Chen, Ian H. Dinwoodie, and Seth Sullivant. Sequential importance sampling for multiway tables. *The Annals of Statistics*, (34):523–545, 2006.

Marco Cuturi, Kenji Fukumizu, and Jean-Philippe Vert. Semigroup kernels on measures. *JMLR*, 6:1169–1198, 2005.

Persi Diaconis and Anil Gangolli. Rectangular arrays with fixed margins. In D. Aldous, P. Diaconis, J. Spencer, and J.M. Steele, editors, *Discrete Probability and Algorithms*, volume 72 of *The IMA Volumes in Mathematics and Its Applications*, pages 15–41. Springer-Verlag, 1995.

Kristen Grauman and Trevor Darrell. Fast contour matching using approximate earth mover's distance. In *IEEE Conf. Vision and Patt. Recog.*, pages 220–227, 2004.

David Haussler. Convolution kernels on discrete structures. Technical report, UC Santa Cruz, 1999. CRL-99-10.

M. Hein and O. Bousquet. Hilbertian metrics and positive definite kernels on probability measures. In *Proceedings of AISTATS*, January 2005.

Mark Jerrum, Alistair Sinclair, and Eric Vigoda. A polynomial-time approximation algorithm for the permanent of a matrix with nonnegative entries. *J. ACM*, 51(4):671–697, 2004.

Thorsten Joachims. *Learning to Classify Text Using Support Vector Machines: Methods, Theory, and Algorithms*. Kluwer Academic Publishers, 2002.

Donald E. Knuth. Permutations, matrices, and generalized Young tableaux. *Pacific J. Math.*, 34:709–727, 1970.

Risi Kondor and Tony Jebara. A kernel between sets of vectors. In T. Faucett and N. Mishra, editors, *Proc. of ICML'03*, pages 361–368, 2003.

John Lafferty and Guy Lebanon. Diffusion kernels on statistical manifolds. *JMLR*, 6:129–163, January 2005.

Guy Lebanon. Metric learning for text documents. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):497–508, 2006.

Jesús A. De Loera, Raymond Hemmecke, Jeremiah Tauzer, and Ruriko Yoshida. Effective lattice point counting in rational convex polytopes. *Journal of Symbolic Computation*, 38(4):1273–1302, October 2004.

Assaf Naor and Gideon Schechtman. Planar earthmover is not in l_1, 2005. arXiv:cs/0509074.

Y. Rubner, C. Tomasi, and L.J. Guibas. The earth mover's distance as a metric for image retrieval. *IJCV: International Journal of Computer Vision*, 40, 2000.

Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization , Optimization, and Beyond*. MIT Press, 2002.

Jean-Philippe Vert, Hiroto Saigo, and Tatsuya Akutsu. Local alignment kernels for protein sequences. In B. Schölkopf, K. Tsuda, and J.-P. Vert, editors, *Kernel Methods in Computational Biology*. MIT Press, 2004.

Cédric Villani. *Topics in Optimal Transportation*, volume 58. AMS Graduate Studies in Mathematics, 2001.