POSITIVITY AND TRANSPORTATION

MARCO CUTURI

Abstract. We prove in this paper that the weighted volume – or generating function – of the set of integral transportation matrices between two integral histograms r and c of equal sum is a positive definite kernel of r and c when the set of considered weights forms a positive definite matrix. The computation of this quantity, despite being the subject of a significant research effort in algebraic statistics, remains an intractable challenge for histograms of even modest dimensions. We propose an alternative kernel which, rather than considering all matrices of the transportation polytope, only focuses on a sub-sample of its vertices known as its Northwestern corner solutions. The resulting kernel is positive definite and can be computed with a number of operations $O(R^2 d)$ that grows linearly in the complexity of the dimension d, where R^2 – the total amount of sampled vertices – is a parameter that controls the complexity of the kernel.

1. Introduction

Suppose that among 30 students in a classroom, 7 and 23 have light and dark colored eyes respectively. You are also told that 12 of them have light hair while 18 have dark hair. What are all the possible populations of the 4 subgroups of students with light/light, dark/dark, light/dark and dark/light eyes and hair color respectively? Such quantities can be arranged in a 2×2 matrix whose row sum vector must be equal to $[7, 23]^T$ and column sum vector must be equal to $[12, 18]$, $\left[\begin{smallmatrix} 3 & 4 \\ 9 & 14 \end{smallmatrix}\right]$ for instance, and more generally *any* integer values in the dots below that satisfy these constraints:

$$\begin{array}{c} \\ 7 \\ 23 \end{array} \begin{array}{c} \begin{array}{cc} 12 & 18 \end{array} \\ \begin{bmatrix} \bullet & \bullet \\ \bullet & \bullet \end{bmatrix} \end{array}$$

Alternatively, suppose that two bakeries in a small village produce daily 7 and 23 loafs of bread each, while two restaurants in the same area each need 12 and 18 loafs to serve their customers every day. What are all the possible morning delivery plans of bread loafs that the two bakeries and shops can agree upon? These seemingly trivial sets of matrices coincide, and are known in the statistics and optimization literature as the sets of *contingency tables* and *transportation plans* respectively.

In statistics, the problem of enumerating all such tables arises naturally in hypothesis testing. Suppose that by entering the aforementioned classroom you observe that the actual repartition of these groups is $\left[\begin{smallmatrix} 5 & 2 \\ 7 & 16 \end{smallmatrix}\right]$. Such an observation intuitively suggests that eye and hair color are related, but how confident should you be about this statement? In the 2×2 case presented above, the Fisher exact test (Yates, 1934) answers that question by computing the probabilities of *all* possible tables outcomes if one assumes that they have been generated as the product

of independent Bernoulli variables with law $p_1 = 7/30$ and $p_2 = 12/30$. By comparing all these probabilities with that of the observed table, we can conclude how reliable an independence hypothesis would be. In optimization, given a 2×2 cost matrix which describes the cost (in gas, calories or time) of bringing a loaf from each bakery to each shop, finding the delivery plan with minimal cost is known as a transportation problem. Transportation problems are an extremely general class of linear programs which are known to encompass all instances of network flows (Bertsimas and Tsitsiklis, 1997, p.274).

Optimal transportation distances (Rachev and Rüschendorf, 1998; Villani, 2009) are distances between probability densities which combine both perspectives outlined above, where the probabilistic view on contingency tables is matched with the goal of computing an optimal transportation plan between two marginal probabilities given a metric on the probability space of interest. Such distances have been widely used in computer vision following the impulsion of Rubner et al. (1997) who used it to compare histograms of image features. When used in information retrieval tasks, transportation distances fare usually better in practice than other classical distances for histograms (Pele and Werman, 2009).

Transportation distances have however two notable drawbacks. First, from a geometric point of view, transportation distances are deficient in the sense that they are not negative definite nor Hilbertian. Negative definiteness carries many favorable properties, among which the possibility to create Euclidean embeddings from which the metric can be accurately recovered, as well as the possibility to turn the distance into a positive definite kernel by simple exponentiation, as a radial basis function. Because of this deficiency, there is no known positive definite counterpart to transportation distances that can leverage the complexity of the set of contingency tables. Second, from a computational point of view, the computational cost of computing transportation distances grows in most cases of interest at least quadratically in the dimension d of the histograms, which can be prohibitive for many applications.

We try to address both issues in this work. The main contribution of this paper is theoretical: after providing some background material and motivation in Section 2 we prove in Section 3 that the generating function of the set of all contingency tables between two integral histograms is a positive definite kernel. Our second contribution is practical: we propose in Section 4 a positive definite kernel that leverages these ideas while still being computationally tractable.

2. Background

2.1. **The Transportation Polytope and the Set of Contingency Tables.** We review in this section a few definitions, notations and results of interest to prove our result. In the following, we write $\langle \cdot, \cdot \rangle$ for both the Frobenius dot-product and the usual dot-product of vectors.

Given a dimension d fixed throughout this paper, for two vectors $r, c \in \mathbb{R}^d$, let $U(r, c)$ be the transportation polytope of r and c, namely the subset of nonnegative matrices in $\mathbb{R}^{d \times d}$ defined as:

$$U(r, c) \stackrel{\text{def}}{=} \{X \in \mathbb{R}_+^{d \times d} \mid X \mathbf{1}_d = r, X^T \mathbf{1}_d = c\},$$

where $\mathbf{1}_d$ is the d dimensional vector of ones. $U(r, c)$ contains all nonnegative $d \times d$ matrices with row and column sums r and c respectively. It is easy to check that

$U(r,c)$ is non-empty if and only if all coordinates of r and c are non-negative and if the total masses of r and c are the same, that is $r^T\mathbf{1}_d = c^T\mathbf{1}_d$. We will consider in most of this work *integral* vectors r and c taken in the set Σ_N of d-dimensional integral histograms with total mass $N \in \mathbb{N}$,

$$\Sigma_d^N \stackrel{\text{def}}{=} \{r \in \mathbb{N}^d \mid r_1 + \cdots + r_d = N\}.$$

We will also focus accordingly on the subset $\mathbb{U}(r,c)$ of $U(r,c)$ that contains all integral transportation matrices, alternatively known as *contingency tables* (Lauritzen, 1982; Everitt, 1992):

$$\mathbb{U}(r,c) \stackrel{\text{def}}{=} U(r,c) \cap \mathbb{N}^{d\times d}.$$

2.2. **Weighted Volumes of Contingency Tables and Particular Cases of Positivity.** Ranging from early work by Yates (1934); Good (1976) to Diaconis and Efron (1985); Cryan and Dyer (2003); Chen et al. (2005), the computation of elementary statistics about $\mathbb{U}(r,c)$ has attracted considerable attention. Many of the ideas of this paper build upon recent work by Barvinok, most notably on his study of the generating function of $\mathbb{U}(r,c)$, defined for $M \in \mathbb{R}^{d\times d}$ as

$$V(r,c\,;M) \stackrel{\text{def}}{=} \sum_{X\in\mathbb{U}(r,c)} e^{-\langle X,M\rangle}.$$

The generating function can be related to the *weighted* volume (Barvinok, 2008, p.2) of $\mathbb{U}(r,c)$, defined for any nonnegative $d\times d$ matrix $K \in \mathbb{R}_+^{d\times d}$ as:

$$T(r,c\,;K) \stackrel{\text{def}}{=} \sum_{X\in\mathbb{U}(r,c)} \prod_{ij}^{d} k_{ij}^{x_{ij}}.$$

Both definitions are equivalent since if we agree that $k_{ij} = e^{-m_{ij}}$ then $T(r,c\,;K) = V(r,c\,;M)$. Because all of our results rely on K's properties, we will mostly use the weighted volume formulation in this paper. Some sections in this paper, notably §2.3 below and §4, are better understood with the generating function formulation.

Cuturi (2007, Prop.2) proved that the cardinal of the set $\mathbb{U}(r,c)$ is a positive definite kernel of r and c using the Robinson-Schensted-Knuth bijection (Knuth, 1970) that maps each contigency table to a pair of Young tableaux with contents r and c and the same pattern. It is easy to see that the cardinal of $\mathbb{U}(r,c)$ is equal to $T(r,c;\mathbf{1}_{d\times d})$ or $V(r,c\,;\mathbf{0}_{d\times d})$. Cuturi (2007, Prop.1) also proved that $T(r,c\,;K)$ is a positive definite kernel of r and c if both are *binary* histograms and K is a nonnegative $d\times d$ positive definite matrix. Since the computation of T entails in that case the computation of the permanent of a Gram matrix, Cuturi (2007) called this kernel the permanent kernel. The main contribution of our paper is to prove in Theorem 1 that the map $(r,c) \in \Sigma_d^N \mapsto T(r,c\,;K)$ is positive definite whenever K is a $d\times d$ positive definite matrix.

2.3. **Relationships with the Optimal Transportation Distance.** Given a $d\times d$ cost matrix M, one can quantify the cost of mapping r to c using a transportation matrix X as $\langle X, M\rangle$. The minimum of this cost is called the optimal transportation cost, defined as:

$$d_M(r,c) \stackrel{\text{def}}{=} \min_{X\in U(r,c)} \langle X, M\rangle.$$

A classical result of optimization in network flows (Bertsimas and Tsitsiklis, 1997, Theo. 7.5) guarantees the existence of a contingency table $X^\star \in \mathbb{U}(r,c)$ which

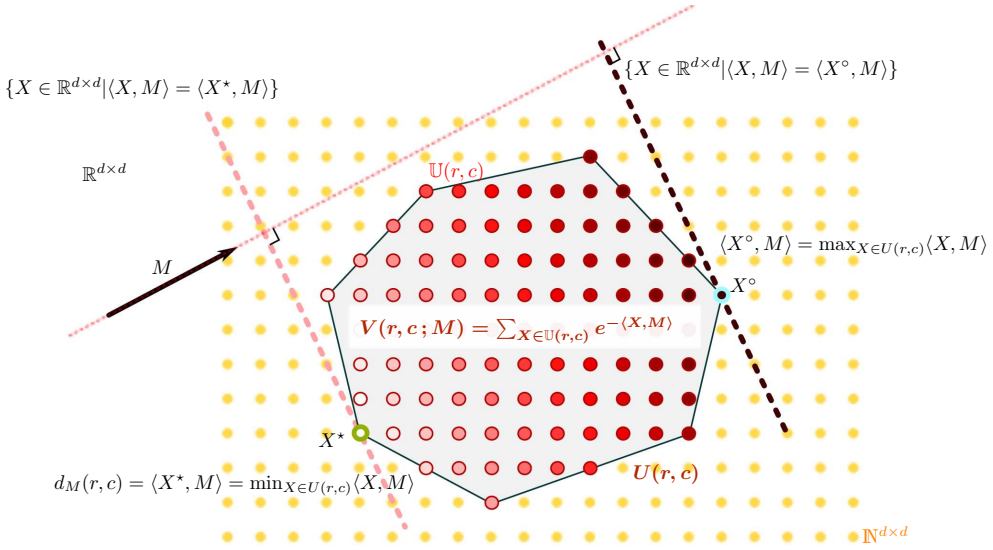

FIGURE 1. Schematic representation of the set $\mathbb{U}(r,c)$ of contingency tables seen as the intersection between the lattice of integral matrices $\mathbb{N}^{d\times d}$ with the transportation polytope $U(r,c)$. Each red dot stands for an integral plan $X \in \mathbb{U}(r,c)$. The inner color in each red dot stands for the value of $\langle X, M\rangle$, which can be seen to go gradually from $\langle X^\star, M\rangle$ to $\langle X^\circ, M\rangle$, that is from the minimum to the maximum of $\langle \cdot, M\rangle$ over $U(r,c)$, or equivalently $\mathbb{U}(r,c)$. The generating function $V(r,c;M)$ of $\mathbb{U}(r,c)$ considers the contributions of *all* contingency tables.

achieves this minimum, as schematically represented in Figure 1. Such an optimal table $X^\star$ can be obtained algorithmically in polynomial time (Ahuja et al., 1993, §9).

The minimal cost $d_M(r,c)$ turns out to be a distance (Villani, 2009, §6.1) whenever the matrix M is itself a metric. This distance is also known as the Wasserstein distance, Monge-Kantorovich's, Mallow's or Earth Mover's (Rubner et al., 1997) in the computer vision literature. The transportation distance is not negative definite in the general case, as shown by counterexamples (Naor and Schechtman, 2007) and embedding distortion results (Andoni et al., 2009). Although some metrics M can yield a negative definite distance[1], characterizing the negative definiteness of d_M remains an open question. Despite this fact, transportation distances have been used in practice to derive a *pseudo*-positive definite kernel: both Jing et al. (2004, §4.C) or Zhang et al. (2006, §2.3) introduce the exponential of (minus) the minimum of $\langle X, M\rangle$,

$$k_M(r,c) = e^{-d_M(r,c)} = \exp\left(-\min_{X\in U(r,c)} \langle X, M\rangle\right), \tag{1}$$

[1]Setting $M = \mathbf{1}_{d\times d} - I_d$ yields the total variation distance between discrete probabilities, which is half the Manhattan or l_1 distance between r and c. All these distances are known to be negative definite.

to form an undefinite kernel which can be used to compare histograms in practice. We prove that, although the value $\exp(-\langle X^\star, M\rangle)$ in itself is not a positive definite kernel, the sum of each term $\exp(-\langle X, M\rangle)$ over *all* possible contingency tables in $\mathbb{U}(r,c)$ is positive definite when M has suitable properties. The generating function V_{rc} can be interpreted as the exponential of (minus) the soft-minimum of $\langle X, M\,\rangle$ over all contingency tables,

$$V(r,c\,;M) = \exp\left(-\operatorname*{softmin}_{X\in\mathbb{U}(r,c)} \langle X, M\,\rangle\right) = e^{\log\sum_{X\in\mathbb{U}(r,c)} e^{-\langle X,M\,\rangle}} = \sum_{X\in\mathbb{U}(r,c)} e^{-\langle X,M\,\rangle},$$

where the soft-minimum of a finite family of scalars (u_i) is

$$\operatorname*{softmin}_i u_i \stackrel{\text{def}}{=} -\log\sum_i e^{-u_i}.$$

This expression relates our results in this work to previous applications of soft minimums to derive positive definite kernels from combinatorial distances for strings (Vert et al., 2004), time series (Cuturi et al., 2007) and trees (Shin et al., 2011). These ideas are summarized in Figure 1.

2.4. **Generalized Permutations.** We close this section by providing some tools to prove the result. We write S_N for the group of permutations over the set $\{1,\cdots,N\}$. For any vector α of size N and permutation $\pi\in S_N$, we write α_π for the permuted vector with coordinates $\alpha_\pi = [\alpha_{\pi(1)}\,\alpha_{\pi(2)}\,\cdots\,\alpha_{\pi(N)}]$ and $\alpha_{p\cdot\cdot q}$ for the subvector $[\alpha_p\,\cdots\,\alpha_q]$ when $1\leq p\leq q\leq N$. For two vectors ρ,γ of $\{1,\cdots,d\}^N$, the $2\times N$ array

$$(\rho\,;\gamma) \stackrel{\text{def}}{=} \begin{bmatrix}\rho_1 & \rho_2 & \cdots & \rho_N\\ \gamma_1 & \gamma_2 & \cdots & \gamma_N\end{bmatrix},$$

is called a generalized permutation (Knuth, 1970). To any generalized permutation $(\rho\,;\gamma)$ corresponds a $d\times d$ integral matrix $\chi(\rho\,;\gamma)$ defined as (Fulton, 1997, p.41):

$$[\chi(\rho\,;\gamma)]_{ij} \stackrel{\text{def}}{=} \sum_{n=1}^{N} \mathbf{1}_{\rho_t=i}\cdot\mathbf{1}_{\gamma_t=j},\quad 1\leq i,j\leq d. \tag{2}$$

Consider the following example where $d=3, N=8$ and

$$\rho = \begin{bmatrix}1\,2\,2\,2\,1\,3\,1\,3\end{bmatrix}, \gamma = \begin{bmatrix}1\,1\,2\,1\,3\,3\,3\,3\end{bmatrix}, (\rho\,;\gamma) = \begin{bmatrix}1\,2\,2\,2\,1\,3\,1\,3\\ 1\,1\,2\,1\,3\,3\,3\,3\end{bmatrix}, \chi(\rho\,;\gamma) = \begin{bmatrix}1 & 0 & 2\\ 2 & 1 & 0\\ 0 & 0 & 2\end{bmatrix}.$$

If we consider now the permutation $\pi = [3\,6\,8\,5\,2\,1\,4\,7]$ we have that

$$\rho = \begin{bmatrix}1\,2\,2\,2\,1\,3\,1\,3\end{bmatrix}, \gamma_\pi = \begin{bmatrix}2\,3\,3\,3\,1\,1\,1\,3\end{bmatrix}, (\rho\,;\gamma_\pi) = \begin{bmatrix}1\,2\,2\,2\,1\,3\,1\,3\\ 2\,3\,3\,3\,1\,1\,1\,3\end{bmatrix}, \chi(\rho\,;\gamma_\pi) = \begin{bmatrix}2 & 1 & 0\\ 0 & 0 & 3\\ 1 & 0 & 1\end{bmatrix}.$$

Note that if ρ and γ have respectively r_i and c_i elements i among their N coefficients for all $1\leq i\leq d$, then $\chi(\rho\,;\gamma)\in\mathbb{U}(r,c)$. One can see above that the corresponding histograms are $r=[3,3,2]$ and $c=[3,1,4]$ and that both $\chi(\rho\,;\gamma)$ and $\chi(\rho\,;\gamma_\pi)$ have row and column sums r and c.

3. The Weighted Volume as a Positive Definite Kernel

Theorem 1. *Let $K \in \mathbb{R}_+^{d\times d}$. The map $(r,c) \mapsto T(r,c\,;K)$ is positive definite if K is positive definite.*

The proof relies on the following observation: Barvinok (2008) showed that the weighted volume of $\mathbf{U}(r,c)$ of two integral histograms r and c of total mass N can be formulated as the expectation of the permanent of a random $N \times N$ matrix. To do so, Barvinok shows that the weighted volume – a sum indexed over all *contigency tables* $X \in \mathbb{U}(r,c)$, can be rewritten as a sum indexed over all *permutations* π in S_N, up to a correcting term known as the Fisher-Yates statistic (Equation (5) in the Appendix). The crux of Barvinok's proof lies in a randomization scheme – using draws from the exponential law – to cancel out the Fisher-Yates statistic. We adopt a similar route to prove the positivity of T, by proving that the inverse of the Fisher-Yates statistic – defined as $\mathbf{k}_2$ below – is itself positive definite to obtain the result.

Proof. Suppose that $K \in \mathbb{R}_+^{d\times d}$ is positive definite and consider two integral histograms r, c in Σ_d^N. We represent r as a N-dimensional vector $\rho \in \{1,\cdots,d\}^N$,

$$\rho \stackrel{\text{def}}{=} [\underbrace{1,\cdots,1}_{r_1 \text{ times}},\underbrace{2,\cdots,2}_{r_2 \text{ times}},\cdots,\underbrace{d,\cdots,d}_{r_d \text{ times}}],$$

and consider the analogous representation γ for c. Let $\mathbf{k}_1$ and $\mathbf{k}_2$ be the following kernels on (ρ,γ):

$$\mathbf{k}_1(\rho,\gamma) = \prod_{t=1}^{N} k(\rho_t,\gamma_t)\ ,\ \text{where } k(i,j) = k_{ij} \text{ for } 1 \le i,j \le d,$$

$$\mathbf{k}_2(\rho,\gamma) = \frac{1}{r_1!\cdots r_d!}\cdot\frac{1}{c_1!\cdots c_d!}\prod_{ij}^{d} x_{ij}!\ ,\ \text{where } X = \chi(\rho\,;\gamma). \quad \text{(see §2.4, Eq. (2))}$$

The kernel $\mathbf{k}_2$ is the inverse of the Fisher-Yates statistic (Equation (5) in the Appendix) associated to an integral transportation table X and its marginals r and c. $\mathbf{k}_1$ is trivially positive definite. The first group of terms of $\mathbf{k}_2$ is trivially positive definite as a product $f(r)f(c)$ where $f(r) = \frac{1}{r_1!\cdots r_d!}$. We prove that the other term, the product of factorials of x_{ij}, is positive definite in Lemma 3 using the proof strategy of a related result provided in Lemma 2. Lemma 4 proves that when a kernel κ on two vectors is symmetric (the definition is provided in the lemma), the sum $\sum_{\pi\in S_N}\kappa(\rho,\gamma_\pi)$ is itself positive definite. We use this result on the product $\kappa(\rho,\gamma) = \mathbf{k}_1(\rho,\gamma)\,\mathbf{k}_2(\rho,\gamma)$ which is trivially symmetric as the product of two symmetric kernels. We then prove in Lemma 5 that

$$\sum_{\pi\in S_N}\kappa(\rho,\gamma_\pi) = T(r,c\,;K).$$

Since the summation over all permutations in the left hand side is positive definite by Lemma 4, we conclude that $T(r,c\,;K)$ is itself a positive definite kernel as the product of two positive definite kernels. ∎

4. Northwestern Kernel

The weighted volume $T(r,c\,;K)$ cannot be computed exactly even for small dimensions d, and approximations (Barvinok, 2008) are currently both too expensive and too loose to be of practical interest in a machine learning context. We adopt in this section an alternative approach, in which we propose to restrict the sum of elementary contributions $\exp(-\langle X, M\rangle)$ to a subset of extreme points of $U(r,c)$ and obtain a kernel whose computational complexity grows linearly in both the dimension d and the size of the sample of extreme points. The main tool for this approach is provided by the Northwestern corner rule to generate a vertex of $U(r,c)$, which we recall in Section 4.1. We define the Northwester kernel in Section 4.2 and prove that it is positive definite. For any matrix $M \in \mathbb{R}^{d\times d}$, we write $M_{\sigma\sigma'}$ for the row and column permuted matrix whose i,j element is $m_{\sigma(i)\sigma'(j)}$.

4.1. **The Northwestern Corner Rule to Generate Vertices of $U(r,c)$.** The Northwestern corner rule is a heuristic that produces a vertex of the polytope $U(r,c)$ in up to $2d$ operations. The rule starts by giving the highest possible value to x_{11}, and at each step when a highest possible value is given to entry x_{ij} it moves on to x_{ij+1} in case x_{ij} filled column j, or x_{i+1j} in case x_{ij} filled row i. The rule proceeds until x_{nn} has received a value. Here is an example of this sequence assuming $r=[2,5,3]$ and $c=[5,1,4]$:

$$\begin{bmatrix}\bullet & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix} \to \begin{bmatrix}2 & 0 & 0\\ \bullet & 0 & 0\\ 0 & 0 & 0\end{bmatrix} \to \begin{bmatrix}2 & 0 & 0\\ 3 & \bullet & 0\\ 0 & 0 & 0\end{bmatrix} \to \begin{bmatrix}2 & 0 & 0\\ 3 & 1 & \bullet\\ 0 & 0 & 0\end{bmatrix} \to \begin{bmatrix}2 & 0 & 0\\ 3 & 1 & 1\\ 0 & 0 & \bullet\end{bmatrix} \to \begin{bmatrix}2 & 0 & 0\\ 3 & 1 & 1\\ 0 & 0 & 3\end{bmatrix}$$

We write $\mathbf{NW}(r,c)$ for the unique Northwestern corner solution that can be obtained through this heuristic. There is, however, a much larger number of Northwestern corner solutions that can be obtained by permuting arbitrarily the order of r and c separately, computing the corresponding Northwestern corner table, and recovering a table of $\mathbb{U}(r,c)$ by inverting again the order of columns and rows. Setting $\sigma=(3,1,2), \sigma'=(3,2,1)$ we have that $r_\sigma=[3,2,5], c_{\sigma'}=[4,1,5]$ and $\sigma^{-1}=(2,3,1), \sigma'=(3,2,1)$. Observe that:

$$\mathbf{NW}(r_\sigma, c'_\sigma) = \begin{bmatrix}3 & 0 & 0\\ 1 & 1 & 0\\ 0 & 0 & 5\end{bmatrix} \in \mathbb{U}(r_\sigma, c_{\sigma'}),\ \mathbf{NW}_{\sigma^{-1}\sigma'^{-1}}(r_\sigma, c_{\sigma'}) = \begin{bmatrix}0 & 1 & 1\\ 5 & 0 & 0\\ 0 & 0 & 3\end{bmatrix} \in \mathbb{U}(r,c).$$

Let $\mathcal{N}(r,c)$ be the set of all Northwestern corner solutions that can be produced this way:

$$\mathcal{N}(r,c) \stackrel{\text{def}}{=} \{\mathbf{NW}_{\sigma^{-1}\sigma'^{-1}}(r_\sigma, c_{\sigma'}), \sigma, \sigma' \in S_d\}.$$

Note that all Northwestern corner solutions only have by construction up to $2d-1$ nonzero elements. The Northwestern corner rule produces a table which is by construction unique for r and c, but there is an exponential number of pairs or row/column permutations (σ,σ') that may share the same table (Stougie, 2002, p.2). $\mathcal{N}(r,c)$ is a subset of the set of extreme points of $U(r,c)$ (Brualdi, 2006, Corollary 8.1.4). $\mathbf{NW}(r,c)$ is an optimal transportation between r and c if the cost matrix M is a Monge matrix (Hoffman, 1961), that is a matrix M that satisfies the inequalities

$$\forall 1 \le i,j,k,l \le d, \quad m_{ij} + m_{kl} \le m_{il} + m_{kj}.$$

Note however that a distance matrix cannot be a Monge matrix since the inequality above applied to $k = j$ and $l = i$ would imply that $0 < 2m_{ij} \leq m_{ii} + m_{jj} = 0$.

4.2. **Random Sampling of Northwestern Corner Solutions.** We propose in this section a kernel which uses arbitrary row/column permutations of r and c to recover extreme points of $\mathbb{U}(r, c)$ and sum their individual contribution:

Theorem 2. *Let R be an arbitrary subset of permutations in S_d. The Northwestern kernel sampled on R and parameterized by a matrix M, defined as*

$$N(r, c\,; K, R) \stackrel{\text{def}}{=} \sum_{\sigma,\sigma' \in R} \exp\left(-\langle M, \mathbf{NW}_{\sigma^{-1}\sigma'^{-1}}(r_\sigma, c_{\sigma'})\rangle\right),$$

is a positive definite kernel if K, the element-wise exponential of $-M$, is positive definite.

Proof. As in the proof of Theorem 1, consider the representation of an integral histogram $r \in \Sigma_d^N$ as a N dimensional vector ρ that replicates r_i times the index i for all i from 1 to d. We also define, for any permutation σ of S_d, the vector ρ_σ as

$$\rho_\sigma \stackrel{\text{def}}{=} [\underbrace{\sigma(1), \cdots, \sigma(1)}_{r_{\sigma(1)} \text{ times}}, \underbrace{\sigma(2), \cdots, \sigma(2)}_{r_{\sigma(2)} \text{ times}}, \cdots, \underbrace{\sigma(d), \cdots, \sigma(d)}_{r_{\sigma(d)} \text{ times}}].$$

ρ_σ for $\sigma \in S_d$ should not be confused with ρ_π for $\pi \in S_N$ (§2.4): for any permutation $\sigma \in S_d$ there exists at least one permutation $\pi \in S_N$ such that $\rho_\sigma = \rho_\pi$ but the converse is not usually true. We show in Lemma 1 that for $\sigma, \sigma' \in S_d$, $\mathbf{NW}_{\sigma^{-1}\sigma'^{-1}}(r_\sigma, c_{\sigma'}) = \chi(\rho_\sigma, \gamma_{\sigma'})$, and thus,

$$N(r, c\,; K, R) = \sum_{\sigma,\sigma' \in R} e^{-\langle M, \chi(\rho_\sigma, \gamma_{\sigma'})\rangle} = \sum_{\sigma,\sigma' \in R} \mathbf{k_1}(\rho_\sigma, \gamma_{\sigma'}),$$

where $\mathbf{k_1}$ is defined in Theorem 1. $N(r, c\,; K, R)$ is positive definite as a convolution kernel. ■

Lemma 1. *Let σ and σ' be two permutations of S_d. Then*

$$\mathbf{NW}_{\sigma^{-1}\sigma'^{-1}}(r_\sigma, c_{\sigma'}) = \chi(\rho_\sigma, \gamma_{\sigma'}).$$

Proof. We write E_{ij} for the $d \times d$ matrix of zeros except for the (i, j) element set to 1. We prove the result by induction on the total mass N. For $N = 1$ the result is trivial since the only transportation matrix in $U(r, c)$ in that case is $E_{\sigma(i_1)\sigma(i_2)}$, where i_1 and i_2 are such that $r_{i_1} = c_{i_2} = 1$. Suppose now that the result is true for all histograms of mass N and consider the case where $r^T\mathbf{1}_d = c^T\mathbf{1}_d = N + 1$. Let i_1 and i_2 be the smallest indices such that $r_{\sigma(i)} > 0$ and $c_{\sigma'(i)} > 0$ respectively. As a consequence, the first elements of ρ_σ and $\gamma_{\sigma'}$ are $\sigma(i_1)$ and $\sigma(i_2)$ respectively. Consider the two vectors ρ_* and γ_* of length N equal to ρ_σ and $\gamma_{\sigma'}$ *without* these two first elements. Setting $\tilde{r}$ and $\tilde{c}$ to r and c except for the fact that $\tilde{r}_{\sigma(i_1)} = r_{\sigma(i_1)} - 1$ and $\tilde{c}_{\sigma(i_2)} = r_{\sigma(i_2)} - 1$, we have by induction that $\mathbf{NW}_{\sigma^{-1}\sigma'^{-1}}(\tilde{r}_\sigma, \tilde{c}_{\sigma'}) = \chi(\rho_*, \gamma_*)$, since the two histograms have total mass N and their representations are respectively ρ_* and γ_*. By definition of the Northwestern corner rule, adding a unit of mass to the i_1's and i_2's components of $\tilde{r}_\sigma$ and $\tilde{c}_{\sigma'}$ only changes the very first iteration of the rule, since all coordinates of $\tilde{r}_\sigma$ and $\tilde{c}_{\sigma'}$ up to but not including i_1 and i_2 respectively are null by construction. Applying the

rule yields a transportation table with an added unit in location (i_1, i_2), providing thus the identity

$$\mathbf{NW}(r_\sigma, c_{\sigma'}) = \mathbf{NW}(\tilde{r}_\sigma, \tilde{c}_{\sigma'}) + E_{i_1 i_2},$$

which implies that

$$\mathbf{NW}_{\sigma^{-1}\sigma'^{-1}}(r_\sigma, c_{\sigma'}) = \mathbf{NW}_{\sigma^{-1}\sigma'^{-1}}(\tilde{r}_\sigma, \tilde{c}_{\sigma'}) + E_{\sigma(i_1)\sigma'(i_2)}. \tag{3}$$

By definition of χ we have that

$$\chi(\rho_\sigma \gamma_\sigma) = \chi(\rho_*, \gamma_*) + E_{\sigma(i_1)\sigma'(i_2)} \tag{4}$$

we get by combining Equations (4) and (3) above with the induction hypothesis that $\mathbf{NW}_{\sigma^{-1}\sigma'^{-1}}(r_\sigma, c_{\sigma'}) = \chi(\rho_\sigma, \gamma_{\sigma'})$. ■

Remark 1. *The evaluation of $N(r, c; K, R)$ requires $O(d|R|^2)$ steps since computing each of the $|R|^2$ contributions $\exp(-\langle M, \mathbf{NW}_{\sigma^{-1}\sigma'^{-1}}(r_\sigma, c_{\sigma'})\rangle)$ for a couple σ, σ' requires up to $2d$ products. The size of $R \subset S_d$ can be controlled from a few permutations to an exhaustive enumeration, which would entail an overall complexity of the order of $O(dd!^2)$.*

5. Conclusion and Future Work

We have proved in this paper that the fundamental ingredient of transportation distances, the polytope of contingency tables, can be used to define a positive definite kernel between two histograms. While the cost matrix of a transportation problem between two histograms r and c needs to be a distance matrix for the optimum to be itself a distance of r and c, we have proved that the generating function of the same polytope is positive definite whenever the cost matrix is itself positive definite. This quantity is computationally intractable, and we have resorted to a summation that only considers a subset of extreme points of the polytope to define the north-western kernel. Future research includes the proposal of suitable subsets R of permutations of S_d tuned with data, as well as other approximation schemes.

Appendix: Intermediate Results for the Proof of Theorem 1

Lemma 2. *Let $a, b \in \{0,1\}^N$ be two binary vectors. The kernel $(a, b) \mapsto \langle a, b\rangle!$ is positive definite.*

Proof. For $N = 1$ the kernel is always equal to 1 and is thus trivially positive definite. For $N > 1$, the recursion $\langle a, b\rangle! = \langle a_1^{N-1}, b_1^{N-1}\rangle!\,(a_N b_N \langle a_1^{N-1}, b_1^{N-1}\rangle + 1)$ provides the expression

$$\langle a, b\rangle! = \prod_{t=1}^{N-1} (a_{t+1} b_{t+1} \langle a_{1..t}, b_{1..t}\rangle + 1),$$

which shows that $\langle a, b\rangle!$ is the product of $N-1$ positive definite kernels on different features of a and b. ■

Remark 2. *Rather than the lemma itself, we will use the identity above in the proof of Lemma 3. We conjecture that this result can be extended to integral vectors. Numerical counterexamples show that this result cannot be generalized to vectors of $\mathbb{R}^N$ through Euler's or Hadamard's Γ function.*

Lemma 3. *Let $\rho, \gamma \in \{1, \cdots, d\}^N$. The kernel $(\rho, \gamma) \mapsto \prod_{ij} x_{ij}!$, where $X = \chi(\rho; \gamma)$, is positive definite.*

Proof. An integral vector $\rho \in \{1, \cdots, d\}^N$ with N components can be represented as a family of d binary row vectors $\rho^1, \cdots, \rho^d$ of length N where for $n \leq N$, $\rho^i_n \stackrel{\text{def}}{=} \mathbf{1}_{\rho_n = i}$. For instance,

$$\text{if } \rho = \begin{bmatrix} 1\,1\,2\,2\,2\,1\,3\,1\,3\,3 \end{bmatrix}, \text{ then } \begin{bmatrix} \rho^1 \\ \rho^2 \\ \rho^3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}$$

These d binary vector representations can be used to obtain the matrix $\chi(\rho\,;\gamma)$. Indeed, it is easy to check that if $X = \chi(\rho\,, \gamma)$ then $x_{ij} = \langle \rho^i, \gamma^j \rangle$. As a consequence, we have that for all indices i, j the coefficient $x_{ij}! = \langle \rho^i, \gamma^j \rangle!$. We obtain that the product of factorials

$$\prod_{ij}^{d} x_{ij}! = \prod_{i,j}^{d} \langle \rho^i, \gamma^j \rangle!,$$

is thus a product of kernels evaluated on all possible pairs among the $d \times d$ representations for ρ and γ. Although one might be tempted to interpret this product as a convolution kernel (Haussler, 1999) or a mapping kernel (Shin and Kuboyama, 2008), one should recall that such results only apply to *sums* of local kernels and not to *products*. Such products of kernels on parts are not, as simple counterexamples can show, positive definite in the general case. Using the decomposition which was used in the proof of Lemma 2, we have however that:

$$\begin{aligned} \prod_{ij}^{d} x_{ij}! = \prod_{i,j}^{d} \langle \rho^i, \gamma^j \rangle! &= \prod_{i,j}^{d} \prod_{t=1}^{N-1} \left(\rho^i_{t+1} \gamma^j_{t+1} \langle \rho^i_{1..t}, \gamma^i_{1..t} \rangle + 1 \right), \\ &= \prod_{t=1}^{N-1} \prod_{i,j}^{d} \left(\rho^i_{t+1} \gamma^j_{t+1} \langle \rho^i_{1..t}, \gamma^j_{1..t} \rangle + 1 \right) = \prod_{t=1}^{N-1} \left(1 + \sum_{i,j}^{d} \rho^i_{t+1} \gamma^j_{t+1} \langle \rho^i_{1..t}, \gamma^j_{1..t} \rangle \right), \end{aligned}$$

where we have used in the last operation the fact that only one of all d^2 products $(\rho^i_{t+1} \gamma^j_{t+1})_{ij}$ is nonzero, since

$$\rho^i_{t+1} \gamma^j_{t+1} = \begin{cases} 1, \text{ if } \rho_{t+1} = i \text{ and } \gamma_{t+1} = j, \\ 0, \text{else.} \end{cases}$$

The product of factorials is thus a product of $N-1$ positive definite kernels indexed by t and defined on ρ and γ, where each of these $N-1$ kernel is 1 plus a convolution kernel operating on the d decompositions of $\rho_{1..t}$ and $\gamma_{1..t}$ as d binary feature vectors, that is

$$\prod_{ij}^{d} x_{ij}! = \prod_{t=1}^{N-1} \left(1 + k_t(\rho, \gamma) \right);$$

where

$$k_t(\rho, \gamma) = \sum_{i,j}^{d} h_t(\rho^i, \gamma^j) \text{ and } h_t(a, b) = a_{t+1} b_{t+1} \langle a_{1..t}, b_{1..t} \rangle.$$

■

Lemma 4. *Let $\alpha = (\alpha_1, \cdots, \alpha_N)$ and $\beta = (\beta_1, \cdots, \beta_N)$ be two lists of N elements in a set $\mathcal{X}$. Let k be a symmetric kernel in $\mathcal{X}^N$, that is a kernel invariant under a permutation of the order of both α and β: $\forall \pi \in S_N,\ k(\alpha, \beta) = k(\alpha_\pi, \beta_\pi)$. Then $(\alpha, \beta) \mapsto \sum_{\pi \in S_N} k(\alpha, \beta_\pi)$ is positive definite.*

Proof. The function g defined below is, by Haussler's (1999) convolution kernels framework, a positive definite kernel of α and β:

$$g(\alpha, \beta) = \sum_{\pi' \in S_N} \sum_{\pi \in S_N} k(\alpha_{\pi'}, \beta_\pi).$$

Using the symmetric property of κ, we have that

$$g(\alpha, \beta) = \sum_{\pi' \in S_N} \sum_{\pi \in S_N} k(\alpha, \beta_{\pi'^{-1} \circ \pi}) = N! \sum_{\pi \in S_N} k(\alpha, \beta_\pi).$$

which proves the result. ■

Lemma 5. $\sum_{\pi \in S_N} \kappa(\rho, \gamma_\pi) = r_1! \cdots r_d! \cdot c_1! \cdots c_d!\, T(r, c\,; K)$

Proof. For any couple of vectors ρ, γ we have that both $\mathbf{k}_1$ and $\mathbf{k}_2$ only depend on $X = \chi(\rho\,;\gamma)$. This is implicitly the case in the definition of $\mathbf{k}_2$ and one can check that

$$\mathbf{k}_1(\rho, \gamma) = \prod_{t=1}^{N} k(\rho_t, \gamma_t) = \prod_{ij}^{d} k_{ij}^{x_{ij}}, \text{ where } X = \chi(\rho\,;\gamma).$$

With every permutation π of we associate a transportation table $\chi(\rho\,;\gamma_\pi)$ which we call the pattern of π. Following (Barvinok, 2008, §2,p.7), we know that the number of permutations π that share the same pattern X for $X \in \mathbb{U}(r, c)$ only depends on X, r and c through a formula known as the Fisher-Yates statistic $n(X)$ of X,

$$n(X) \stackrel{\text{def}}{=} \text{card}\{\pi \in S_N |\, \chi(\rho\,;\gamma_\pi) = X\} = \frac{r_1! \cdots r_d! \cdot c_1! \cdots c_d!}{\prod_{ij} x_{ij}!}. \tag{5}$$

We thus have that

$$\begin{aligned} \sum_{\pi \in S_N} \kappa(\rho, \gamma_\pi) &= \sum_{X \in \mathbb{U}(r,c)} n(X)\, \mathbf{k}_1(\rho, \gamma_\pi) \mathbf{k}_2(\rho, \gamma_\pi) \\ &= \sum_{X \in \mathbb{U}(r,c)} \frac{r_1! \cdots r_d! \cdot c_1! \cdots c_d!}{\prod_{ij}^{d} x_{ij}!} \prod_{ij}^{d} k_{ij}^{x_{ij}} \frac{\prod_{ij}^{d} x_{ij}!}{r_1! \cdots r_d! \cdot c_1! \cdots c_d!} = T(r, c\,; K). \end{aligned}$$

■

References

Ahuja, R., Magnanti, T., and Orlin, J. (1993). *Network Flows: Theory, Algorithms and Applications*. Prentice Hall.

Andoni, A., Ba, K. D., Indyk, P., and Woodruff, D. (2009). Efficient sketches for earth-mover distance, with applications. In *Foundations of Computer Science (FOCS) 2009.*, pages 324 –330.

Barvinok, A. (2008). Enumerating contingency tables via random permanents. *Combinatorics, Probability and Computing*, 17(1):1–19.

Bertsimas, D. and Tsitsiklis, J. (1997). *Introduction to linear optimization*. Athena Scientific.

Brualdi, R. (2006). *Combinatorial matrix classes*. Encyclopedia of Mathematics and Its Applications 108, Cambridge University Press.

Chen, Y., Diaconis, P., Holmes, S., and Liu, J. (2005). Sequential monte carlo methods for statistical analysis of tables. *Journal of the American Statistical Association*, 100(469):109–120.

Cryan, M. and Dyer, M. (2003). A polynomial-time algorithm to approximately count contingency tables when the number of rows is constant. *Journal of Computer and System Sciences*, 67(2):291–310.

Cuturi, M. (2007). Permanents, transportation polytopes and positive-definite kernels on histograms. In *Proc. of the 20th Intern. Joint Conf. on Artificial Intelligence 2007*, pages 732 – 737.

Cuturi, M., Vert, J.-P., Birkenes, Ø.., and Matsui, T. (2007). A kernel for time series based on global alignments. In *Proceedings of ICASSP*, volume II, pages 413 – 416.

Diaconis, P. and Efron, B. (1985). Testing for independence in a two-way table: new interpretations of the chi-square statistic. *The Annals of Statistics*, 13(3):845–913.

Everitt, B. (1992). *The analysis of contingency tables*. Chapman & Hall/CRC.

Fulton, W. (1997). *Young tableaux: with applications to representation theory and geometry*, volume 35. Cambridge Univ Press.

Good, I. J. (1976). On the application of symmetric dirichlet distributions and their mixtures to contingency tables. *The Annals of Statistics*, 4(6):pp. 1159–1189.

Haussler, D. (1999). Convolution kernels on discrete structures. Technical report, UCSC. USCS-CRL-99-10.

Hoffman, A. (1961). On simple linear programming problems. In *Proceedings of Symposia in Pure Mathematics*, volume 7, pages 317–327. American Mathematical Society.

Jing, F., Li, M., Zhang, H.-J., and Zhang, B. (2004). An efficient and effective region-based image retrieval framework. *Image Processing, IEEE Transactions on*, 13(5):699 –709.

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

Lauritzen, S. (1982). *Lectures on contingency tables*. Aalborg Univ. Press.

Naor, A. and Schechtman, G. (2007). Planar earthmover is not in l_1. *SIAM J. Comput.*, 37(3):804–826.

Pele, O. and Werman, M. (2009). Fast and robust earth mover's distances. In *ICCV'09*.

Rachev, S. and Rüschendorf, L. (1998). *Mass Transportation Problems: Theory*, volume 1. Springer Verlag.

Rubner, Y., Guibas, L., and Tomasi, C. (1997). The earth movers distance, multi-dimensional scaling, and color-based image retrieval. In *Proceedings of the ARPA Image Understanding Workshop*, pages 661–668.

Shin, K., Cuturi, M., and Kuboyama, T. (2011). Mapping kernels for trees. *Proc. of ICML 2011*.

Shin, K. and Kuboyama, T. (2008). A generalization of Haussler's convolution kernel: mapping kernel. In *Proceedings of the 25th international conference on Machine learning*, pages 944–951.

Stougie, L. (2002). A polynomial bound on the diameter of the transportation polytope. Technical report.

Vert, J.-P., Saigo, H., and Akutsu, T. (2004). Local alignment kernels for protein sequences. In Schölkopf, B., Tsuda, K., and Vert, J.-P., editors, *Kernel Methods in Computational Biology*. MIT Press.

Villani, C. (2009). *Optimal transport: old and new*, volume 338. Springer Verlag.

Yates, F. (1934). Contingency tables involving small numbers and the χ 2 test. *Supplement to the Journal of the Royal Statistical Society*, 1(2):217–235.

Zhang, J., Marszalek, M., Lazebnik, S., and Schmid, C. (2006). Local features and kernels for classification of texture and object categories: A comprehensive study. In *CVPRW '06*, page 13.

Graduate School of Informatics, Kyoto University
E-mail address: mcuturi@i.kyoto-u.ac.jp