The generating function of the polytope of transport matrices $U(r, c)$ as a positive semidefinite kernel of the marginals r and c

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

June 23, 2006

Abstract

The Monge-Kantorovich distance is a popular distance for probability distributions, notably histograms. Given two histograms, or marginals, $r = (r_i)_{1 \leq i \leq d}$ and $c = (c_i)_{1 \leq i \leq d}$ of equal sum N, the Monge-Kantorovich distance $d_{\mathrm{MK}}(r, c)$ between r and c parameterized by a $n \times n$ cost matrix T is the minimum of all costs $< F, T >$ taken over matrices F of the transportation polytope $U(r, c)$. This minimum is reached in $F^\star$, which is known to be a vertex of the polytope and which can be computed through linear programming. Recent results suggest that this distance is not negative definite, and hence $\exp(-\frac{1}{t} d_{\mathrm{MK}})$ may not be a positive definite kernel for many choices of $t > 0$. We prove in this report that the generating function of the subset of integral matrices of $U(r, c)$ evaluated on a fixed cost matrix T is a positive semidefinite function of r and c when T has adequate properties, notably when T is symmetric negative semidefinite matrix. We obtain the result by considering first the simple case where r and c are clouds of points and introducing the permanent of the Gram matrix between elements of r and c as a positive semidefinite kernel. We then follow by adapting an expression of the generating function of the polytope in terms of random permanents recently obtained by A. Barvinok, through a recent result on Stieltjes moments obtained by C.Berg and A.J. Duran in a different context.

Defining meaningful kernels on positive measures, histograms or clouds of points is an important topic in the field of kernel methods. This topic is directly related to the definition of kernels for structured objects seen as bags-of-components, and hence of chief interest for practitioners.

In the early applications of kernel methods to structured data, histograms were often treated as simple vectors and used as such with the standard Gaussian or polynomial kernels [Joa02]. More adequate positive definite[1]kernels

[1]In this paper, we do not distinguish positive semidefinite kernels (resp. matrices) from

which exploit their specificity have been proposed since. Namely, kernels which take into account the fact that histograms are vectors with nonnegative coordinates [HB05], and whose sum may be normalized to one, that is cast as discrete probability measures and treated under the light of information geometry [AN01] as in [LL05, Leb06]. 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 [KJ03, CFV05, HB05].

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 [Vil01], which is usually known as the Monge-Kantorovich (MK) or Wasserstein distance. This distance is also popular in the computer vision community [RTG00] under the name of the earth movers' distance. However, preliminary findings [NS05] suggest that the MK distance is not negative definite, and cannot thus be used directly to define positive definite kernels, through negative exponentiation for instance. Although some approximations [IT03] of the distance have been used so far to define positive definite kernels in vision applications [GD04], 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 methodology used in [VSA04] to define a positive definite kernel for strings derived from sequences of string manipulations which may map a string m_1 to m_2(namely deletion, substitution and insertions of letters). In [VSA04], the authors consider for two strings m_1, m_2 all possible sequences σ of such transformations such that $\sigma(m_1) = m_2$. Each of these admissible sequences σ is weighted by a score criterion $\pi(\sigma)$ which measures how costly the sequence of operations is, to obtain a positive definite kernel between sequences defined as $\sum_\sigma \exp(\beta\pi(\sigma))$ with $\beta > 0$. The obtained kernel performs experimentally better than the naive use of the score of the optimal alignment $\pi(\sigma^\star)$, known as the Smith-Waterman score in the context of biological sequences.

In our setting, the set of possible transformations is played by the set of transportation matrices $F \in U(r, c)$ between two discrete histograms, where the associated cost is $< F, T >$. We name the kernel presented in this paper the Monge-Kantorovich kernel between histograms or marginals, as a reference to the original distance. As is the case in [VSA04], the family of convolution kernels [Hau99] plays an important role in our proofs. We use a key result from [Bar05] to prove the positive definiteness of the kernel, which we slightly adapt by using another recent result obtained by [BD04]. We believe the Monge-Kantorovich kernels might be useful in machine learning tasks, although we expect their computational cost to be problematic at the moment. Exact computations are currently feasible with histograms of only very small dimensions, roughly 5, through different implementations (see [LHTY04] for

positive definite kernels (resp. matrices). Hence *definite* should be always be interpreted as *semi*definite. The same remark applies for negative (semi)definiteness.

instance) of Barvinok's algorithm to compute generating function of polytopes. Therefore, only approximate calculations are possible for larger histograms at the moment. Recent advances in the computation of the permanent through sequential monte-carlo techniques [JSV04] should enable us to compare thousands of clouds of tens of points in a arbitrary space endowed with a kernel, using the formula in Lemma 4. Monge-Kantorovich kernels for histograms might also be computed through SMC approximations of the generating function through recent algorithms [CDS06] which may seemingly compute efficiently kernel values for histograms with tens of bins.

The Monge-Kantorovich kernel

We write $\mathbb{N} = \{0, 1, \ldots\}$ for the set of nonnegative integers. We restrict our study to the case of integral histograms (or margins as in the corresponding statistical literature [DE85]) with the same sum and dimensions, that is elements of the simplex lattice

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

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

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

where $U(r, c) = \{F \in \mathbb{R}_{d \times d} \,|\, F_{ij} \geq 0, F\mathbb{1}_d = r, F^\top \mathbb{1}_d = c\}$ is the polytope of transport matrices between r and c, $T \in \mathbb{R}_{d,d}$ is an arbitrary distance matrix between bins, $\mathbb{1}_d$ is the d-dimensional vector of ones, and for two square matrices U and V we use the Frobenius dot-product $< U, V >= \mathrm{tr}(UV^\top)$.

The optimal plan

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

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 with integral coordinates [KW68]. Rather than only using this optimal vertex to define a kernel between r and c through $\exp(- < F^\star, T >)$, which would not necessarily be positive definite, we propose to use the criterion $< F, T >$ over all integral transportation plans $F \in U(r, c)$ in the following way,

Theorem 1 *The Monge-Kantorovich kernel $k_{MK} : \Sigma_{d,N} \times \Sigma_{d,N} \to \mathbb{R}$, such that*

$$k_{MK}^T(r, c) = \sum_{F \in U(r,c) \cap \mathbb{N}_{d \times d}} e^{-<T,F>},$$

is positive definite if $[e^{-t_{ij}}]_{1 \leq i,j \leq d}$ is a positive semidefinite matrix.

Note that k_{MK}^T is exactly the generating function of the polytope $U(r,c)$ restricted to its lattice points in $\mathbb{N}_{d\times d}$, evaluated in T as defined for instance in [LHTY04]. It can also be considered as the total weight of $U(r,c)$ if we take the notations of [Bar06], by setting $w_{ij} = e^{-t_{ij}}$. Note that the conditions of Theorem 1 can be loosened by asking T to be simply negative definite[2], which is a sufficient condition for $[e^{-t_{ij}}]$ to be positive definite as stated in Schoenberg's theorem [BCR84, Theorem 3.2.2].

The symmetry of k_{MK}^T is ensured by the symmetry of T, since $U(c,r) = U(r,c)^\top$ and we have that $< F, T >=< F^\top, T >$. We prove the positive-definiteness of k_{MK}^T using four lemmas. Recall that for a $n\times n$ matrix $M = [m_{ij}]$, the permanent of M, $\operatorname{per} M$ is defined as

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

where the set S_n is the set of all permutations of $\{1,\ldots,n\}$.

Lemma 2 *Let $\mathcal{X}$ be a set endowed with a kernel κ and $\mathcal{X}^n$ the set of sequences 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 $y = (y_1,\ldots,y_n)$ be two sequences of $\mathcal{X}^n$. Then*

$$k_{\mathrm{per}} : (x,y) \mapsto \operatorname{per}([\kappa(x_i,y_j)]_{1\le i,j\le n})$$

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

Proof. For a sequence x and a permutation σ we write x_σ for the sequence $(x_{\sigma(1)},\ldots,x_{\sigma(n)})$. The proof now follows from Haussler's convolution kernels [Hau99]. Namely, we consider the equivalence relation $\mathcal{R}$ between two sequences x and y of $\mathcal{X}^n$, where $x\mathcal{R}y$ if and only if there exists a permutation $\sigma \in S_n$ such that $x_\sigma = y$. Consider now the kernel k for two sequences

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

which we use to define the convolution kernel

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

hence the positive definiteness of k_{per}.■

We will use the following lemma to adapt the randomized setting proposed by [Bar05] to express the generating function of $U(r,c)$:

[2]Following the terminology of [BCR84], a $d\times d$ real matrix T is negative semidefinite if for all vectors $u\in\mathbb{R}^d$ such that $\sum_{i=1}^d u_i = 0$, $u^\top T u \le 0$. Note that this is not equivalent to stating that $-T$ is positive semidefinite, which is only a sufficient condition.

Lemma 3 (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 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 [BD04] for a proof of this result, and more generally to the reference [BCR84] for the exposition of the moment problems and their relationship with harmonic analysis on semigroups. We will only use in practice $\mu_{\frac{1}{2}}$, and start using the framework of [Bar05] in the following lemma.

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 $\gamma = (\gamma_1, \ldots, \gamma_N)$ 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

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

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 that

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

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

Lemma 5 (Barvinok) *Suppose $\gamma = (\gamma_1, \ldots, \gamma_N)$ is distributed as a sequence of independent random variables with identical law $\mu_{\frac{1}{2}}$. Through the identity*

$$k^T_{MK}(r, c) = E[k_{\gamma,T}(r, c)]$$

we have that k^T_{MK} is positive semidefinite.

$E[k_{\gamma,T}(r, c)]$ is trivially positive definite as a sum of positive definite kernels. We follow Barvinok's proof to prove the identity, with a slight modification: Barvinok considers standard exponential variable γ_{ij} arranged in a $N \times N$ matrix, while we consider here a sequence of independent random variables $\gamma = (\gamma_1, \ldots, \gamma_N)$ which all follow law $\mu_{\frac{1}{2}}$.

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 per A. Hence

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

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 $D = [d_{ij}]_{1 \le i,j \le d}$ where

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

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 $D : \sigma \mapsto D(\sigma) \in U(r,c) \cap \mathbb{N}_{d,d}$ is surjective, but not bijective as we see below. For h_σ, we thus have, through Lemma 3 that

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

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

$$\frac{r_1! \cdots r_d! c_1! \cdots c_d!}{\prod_{i,j} d_{ij}!},$$

yielding

$$E[\text{per } A] = \sum_{D \in U(r,c) \cap \mathbb{N}_{d,d}} \frac{r_1! \cdots r_d! c_1! \cdots c_d!}{\prod_{i,j} d_{ij}!} \times \prod_{i,j} e^{-t_{ij} d_{ij}} d_{ij}!,$$

and hence

$$\begin{aligned} E[k_{\gamma,T}(r,c)] &= \sum_{D \in U(r,c) \cap \mathbb{N}_{d,d}} \prod_{i,j} e^{-t_{ij} d_{ij}} \\ &= \sum_{D \in U(r,c) \cap \mathbb{N}_{d,d}} e^{-<T,D>} = k_{\text{MK}}^T(r,c). \end{aligned}$$

■

References

[AN01] Shun-Ichi Amari and Hiroshi Nagaoka. *Methods of Information Geometry*. AMS vol. 191, 2001.

[Bar05] Alexander Barvinok. Enumerating contingency tables via random permanents, 2005. arXiv.org:math/0511596.

[Bar06] 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.

[BCR84] Christian Berg, Jens Peter Reus Christensen, and Paul Ressel. *Harmonic Analysis on Semigroups*. Number 100 in GTM. Springer Verlag, 1984.

[BD04] Christian Berg and Antonio J. Duran. A transformation from hausdorff to stieltjes moment sequences. *Arkiv för matematik*, 42:239–257, 2004.

[CDS06] Yuguo Chen, Ian H. Dinwoodie, and Seth Sullivant. Sequential importance sampling for multiway tables, 2006.

[CFV05] Marco Cuturi, Kenji Fukumizu, and Jean-Philippe Vert. Semigroup kernels on measures. *JMLR*, 6:1169–1198, 2005.

[DE85] Persi Diaconis and Bradley Efron. Testing for independence in a two-way table: new interpretations of the chi-square statistic. *Ann. Statist.*, 13(3):845–913, 1985.

[GD04] Kristen Grauman and Trevor Darrell. Fast contour matching using approximate earth mover's distance. In *CVPR (1)*, pages 220–227, 2004.

[Hau99] David Haussler. Convolution kernels on discrete structures. Technical report, UC Santa Cruz, 1999. USCS-CRL-99-10.

[HB05] M. Hein and O. Bousquet. Hilbertian metrics and positive definite kernels on probability measures. In Z. Ghahramani and R. Cowell, editors, *Proceedings of AISTATS 2005*, January 2005.

[IT03] Piotr Indyk and Nitin Thaper. Fast image retrieval via embeddings. In *3rd International Workshop on Statistical and Computational Theories of Vision (at ICCV)*, 2003.

[Joa02] Thorsten Joachims. *Learning to Classify Text Using Support Vector Machines: Methods, Theory, and Algorithms*. Kluwer Academic Publishers, 2002.

[JSV04] 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.

[KJ03] 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.

[KW68] V. Klee and C. Witzgall. Facets and vertices of transportation polytopes. In G. B. Dantzig and A. F. Veinott, editors, *Mathematics of the Decision Sciences, Part 1*, pages 257–282, Providence, 1968. American Mathematical Society.

[Leb06] Guy Lebanon. Metric learning for text documents. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):497–508, 2006.

[LHTY04] 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.

[LL05] John Lafferty and Guy Lebanon. Diffusion kernels on statistical manifolds. *JMLR*, 6:129–163, January 2005.

[NS05] Assaf Naor and Gideon Schechtman. Planar earthmover is not in l_1, 2005. arXiv:cs/0509074.

[RTG00] 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.

[Vil01] Cédric Villani. *Topics in Optimal Transportation*, volume 58. AMS Graduate Studies in Mathematics, 2001.

[VSA04] Jean-Philippe Vert, Hiroto Saigo, and Tatsuya Akutsu. Local alignment kernels for protein sequences. In Bernhard Schölkopf, Koji Tsuda, and Jean-Philippe Vert, editors, *Kernel Methods in Computational Biology*. MIT Press, 2004.