Kantorovich Metric in Computer Science: A Brief Survey

Yuxin Deng [a], Wenjie Du [b]

[a] *Shanghai Jiao Tong University, China*
yuxindeng@sjtu.edu.cn

[b] *Shanghai Normal University, China*
wenjiedu@shnu.edu.cn

Abstract

In contrast to its wealth of applications in mathematics, the Kantorovich metric started to be noticed in computer science only in recent years. We give a brief survey of its applications in probabilistic concurrency, image retrieval, data mining, and bioinformatics. This paper highlights the usefulness of the Kantorovich metric as a general mathematical tool for solving various kinds of problems in rather unrelated domains.

Key words: Kantorovich metric, probabilistic concurrency, information retrieval, bioinformatics.

1 Introduction

The *transportation problem* has been playing an important role in linear programming due to its general formulation and methods of solution. The original transportation problem, formulated by the French mathematician G. Monge in 1781 [21], consists of finding an optimal way of shovelling a pile of sand into a hole of the same volume. In the 1940s, the Russian mathematician and economist L.V. Kantorovich, who was awarded a Nobel prize in economics in 1975 for the theory of optimal allocation of resources, gave a relaxed formulation of the problem and proposed a variational principle for solving the problem [16]. Unfortunately, Kantorovich's work went unrecognized during a long period of time. The later known *Kantorovich metric* has appeared in the literature under different names, because it has been rediscovered historically several times from different perspectives. Many metrics known in measure theory, ergodic theory, functional analysis, statistics, etc. are special cases of the general definition of the Kantorovich metric [34]. The elegance of the formulation, the fundamental character of the optimality criterion, as well as the wealth of applications, which keep arising, place the Kantorovich metric in a prominent position among the mathematical works of the 20th century. In addition, this formulation can be computed in polynomial time [22], which is an appealing feature for its use in solving applied problems. For example, it is widely used to solve a variety of problems in business and economy such as market distribution, plant location, scheduling problems etc. However, as far as we know the metric attracted the attention of computer scientists only in recent years. In this short paper, we give a brief survey of recent applications of the Kantorovich metric in computer science. In order to give the reader a feel for how the metric has been used, we take five examples from four different areas in computer science: probabilistic concurrency, image retrieval, data mining, and bioinformatics. In some cases, the metric is directly used to compute similarities of the objects we are interested in; in the other cases, variations of the metric are adapted to meet our requirements. The purpose of this paper is to review some existing applications so as to highlight the usefulness of the Kantorovich metric as a general mathematical tool for solving various kinds of problems in rather unrelated domains.

2 Kantorovich metric

Roughly speaking, the Kantorovich metric provides a way of measuring the distance between two distributions. Of course, this requires first a notion of distance between the basic features that are aggregated into the distributions, which is often referred to as the *ground distance*. For example, in the case of color, the ground distance measures dissimilarity between individual colors. In other words, the Kantorovich metric defines a "lifted" distance, or dissimilarity, between two distributions of mass in a space that is itself endowed with a ground distance. For color, this means finding distances between image color distributions. There are a host of metrics available in the literature (see e.g. [13]) to quantify the distance between probability measures; see [24] for a comprehensive review of metrics in the space of probability measures. The Kantorovich metric has an elegant formulation and a natural interpretation in terms of the transportation problem.

We now recall the mathematical definition of the Kantorovich metric. Let (S, d) be a separable metric space. (This condition will be used by Theorem 2.4 below.)

Definition 2.1 Given any two Borel probability measures $\mathbb{P}$ and $\mathbb{Q}$ on S, the *Kantorovich distance* between $\mathbb{P}$ and $\mathbb{Q}$ is defined by

$$K(\mathbb{P}, \mathbb{Q}) = \sup \left\{ \left| \int f d\mathbb{P} - \int f d\mathbb{Q} \right| : ||f|| \leq 1 \right\}.$$

where $|| \cdot ||$ is the *Lipschitz semi-norm* defined by $||f|| = \sup_{x \neq y} \frac{|f(x)-f(y)|}{d(x,y)}$ for a function $f : S \to \mathcal{R}$ with $\mathcal{R}$ being the set of all real numbers.

The Kantorovich metric has an alternative characterisation. We denote by

$\mathcal{P}(S)$ the set of all Borel probability measures on S such that for all $z \in S$, if $\mathbb{P} \in \mathcal{P}(S)$ then $\int_S d(x,z)\mathbb{P}(x) < \infty$. We write $M(\mathbb{P},\mathbb{Q})$ for the set of all Borel probability measures on the product space $S \times S$ with marginal measures $\mathbb{P}$ and $\mathbb{Q}$, i.e. if $\mu \in M(\mathbb{P},\mathbb{Q})$ then $\int_{y\in S} d\mu(x,y) = d\mathbb{P}(x)$ and $\int_{x\in S} d\mu(x,y) = d\mathbb{Q}(y)$ hold.

Definition 2.2 For $\mathbb{P},\mathbb{Q} \in \mathcal{P}(S)$, we define the metric L as follows:

$$L(\mathbb{P},\mathbb{Q}) = \inf\left\{\int d(x,y)d\mu(x,y) : \mu \in M(\mathbb{P},\mathbb{Q})\right\}.$$

Lemma 2.3 *If (S,d) is a separable metric space then K and L are metrics on $\mathcal{P}(S)$.*

The famous Kantorovich-Rubinstein duality theorem gives a dual representation of K in terms of L.

Theorem 2.4 (Kantorovich-Rubinstein [17]) *If (S,d) is a separable metric space then for any two distributions $\mathbb{P},\mathbb{Q} \in \mathcal{P}(S)$ we have $K(\mathbb{P},\mathbb{Q}) = L(\mathbb{P},\mathbb{Q})$.*

In view of the above theorem, many papers in the literature directly take Definition 2.2 as the definition of the Kantorovich metric. Here we keep the original definition, but it is helpful to understand K by using L. Intuitively, a probability measure $\mu \in M(\mathbb{P},\mathbb{Q})$ can be understood as a *transportation* from one unit mass distribution $\mathbb{P}$ to another unit mass distribution $\mathbb{Q}$. If the distance $d(x,y)$ represents the cost of moving one unit of mass from location x to location y then the Kantorovich distance gives the optimal total cost of transporting the mass of $\mathbb{P}$ to $\mathbb{Q}$. We refer the reader to Villani's book [35] for an excellent exposition on the Kantorovich metric and the duality theorem.

Many problems in computer science only involve finite state spaces, so discrete distributions with finite supports are sometimes more interesting than continuous distributions. For two discrete distributions $\mathbb{P}$ and $\mathbb{Q}$ with finite supports $\{x_1,...,x_n\}$ and $\{y_1,...,y_m\}$, respectively, minimizing the total cost of a discretized version of the transportation problem reduces to the following linear programming problem:

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^n \sum_{j=1}^m \mu(x_i,y_j)d(x_i,y_j) \\ \text{subject to} & \bullet\ \forall 1 \le i \le n : \sum_{j=1}^m \mu(x_i,y_j) = \mathbb{P}(x_i) \\ & \bullet\ \forall 1 \le j \le m : \sum_{i=1}^n \mu(x_i,y_j) = \mathbb{Q}(x_j) \\ & \bullet\ \forall 1 \le i \le n, 1 \le j \le m : \mu(x_i,y_j) \ge 0. \end{array} \tag{1}$$

Since (1) is a special case of the discrete mass transportation problem, some well-known polynomial time algorithm like [22] can be employed to solve it, which is an attractive feature for computer scientists.

3 Applications in probabilistic concurrency

In this section we present two applications of the Kantorovich in probabilistic concurrency: one is in defining behavioural pseudometrics and the other in defining (bi-)simulations.

3.1 Behavioural pseudometrics

The Kantorovich metric has been used by van Breugel *et al.* for defining behavioural pseudometrics on fully probabilistic systems [30,33,29] and reactive probabilistic systems [31,32,27,28]; and by Desharnais *et al.* for labelled Markov chains [8,10] and labelled concurrent Markov chains [9]; and later on by Ferns *et al.* for Markov decision processes [11,12]; and by Deng *et al.* for action-labelled quantitative transition systems [4]. Given a pseudometric m on a finite set of states, a typical problem to measure similarities of the behaviour of probabilistic processes is how to lift m to be a pseudometric $\hat{m}$ on distributions over states [1]. An elegant definition of lifting is via the Kantorovich metric. We illustrate the basic idea by considering the simple case of dealing with distributions over a finite state space.

Let $\mathbb{P}$ and $\mathbb{Q}$ be distributions on a finite set S of states. Suppose the distance $m(s,t)$ between any two states s and t is bounded by 1. In [30] the distance $\hat{m}(\mathbb{P},\mathbb{Q})$ is given by the value of the following linear programming problem:

$$\begin{array}{ll} \text{maximize} & \sum_{s\in S}(\mathbb{P}(s)-\mathbb{Q}(s))x_s \\ \text{subject to} & \bullet\ \forall s,t\in S: x_s - x_t \leq m(s,t) \\ & \bullet\ \forall s\in S: 0\leq x_s\leq 1. \end{array} \tag{2}$$

This problem can be dualized and then simplified to yield the following problem:

$$\begin{array}{ll} \text{minimize} & \sum_{s,t\in S} y_{st}m(s,t) \\ \text{subject to} & \bullet\ \forall s\in S: \sum_{t\in S} y_{st} = \mathbb{P}(s) \\ & \bullet\ \forall t\in S: \sum_{s\in S} y_{st} = \mathbb{Q}(t) \\ & \bullet\ \forall s,t\in S: y_{st}\geq 0. \end{array} \tag{3}$$

Now (3) is in the same form as (1).

3.2 (Bi-)simulations

Given a state space S, a probabilistic bisimulation or simulation is a relation R over states. However, in many models a step of transition leads a state to

[1] To some extent, this is related to measuring the distances between quantum states, so it is reasonable to expect applications of the Kantorovich metric in quantum mechanics as well [37].

a distribution and so when defining (bi-)simulations we need to lift R to be a relation $R^\dagger$ over distributions. One way of lifting [15] is given as follows; there are other equivalent definitions (see e.g. [18,5,7,6]).

Definition 3.1 Given a relation $R \subseteq S \times S$, we lift it to a relation $R^\dagger \subseteq \mathcal{P}(S) \times \mathcal{P}(S)$ by letting $\mathbb{P}\ R^\dagger\ \mathbb{Q}$ whenever there exists a weight function $w : S \times S \to [0,1]$ such that

(i) $\forall s \in S : \sum_{t\in S} w(s,t) = \mathbb{P}(s)$
(ii) $\forall t \in S : \sum_{s\in S} w(s,t) = \mathbb{Q}(t)$
(iii) $\forall s,t \in S : w(s,t) > 0 \Rightarrow s\ R\ t$.

This way of lifting binary relations has an intrinsic connection with the lifting of pseudometrics via the Kantorovich metric given in (3), as stated by the next proposition which is a novel result of the paper.

Proposition 3.2 *Let R be a binary relation and m a pseudometric on a state space S satisfying*

$$s\ R\ t \quad \textit{iff} \quad m(s,t) = 0 \tag{4}$$

for any $s,t \in S$. Then it holds that

$$\mathbb{P}\ R^\dagger\ \mathbb{Q} \qquad \textit{iff} \qquad \hat{m}(\mathbb{P},\mathbb{Q}) = 0$$

for any distributions $\mathbb{P},\mathbb{Q} \in \mathcal{P}(S)$.

Proof. Suppose $\mathbb{P}\ R^\dagger\ \mathbb{Q}$. From Definition 3.1 we know there is a weight function w such that

(i) $\forall s \in S : \sum_{t\in S} w(s,t) = \mathbb{P}(s)$
(ii) $\forall t \in S : \sum_{s\in S} w(s,t) = \mathbb{Q}(t)$
(iii) $\forall s,t \in S : w(s,t) > 0 \Rightarrow s\ R\ t$.

By substituting $w(s,t)$ for $y_{s,t}$ in (3), the three constraints there can be satisfied. For any $s,t \in S$ we distinguish two cases:

(i) either $w(s,t) = 0$
(ii) or $w(s,t) > 0$. In this case we have $s\ R\ t$, which implies $m(s,t) = 0$ by (4).

Therefore, we always have $w(s,t)m(s,t) = 0$ for any $s,t \in S$. Consequently, $\sum_{s,t\in S} w(s,t)m(s,t) = 0$ and the optimal value of the problem in (3) must be 0, i.e. $\hat{m}(\mathbb{P},\mathbb{Q}) = 0$, and the optimal solution is determined by w.

The above reasoning can be reversed to show that the optimal solution of (3) determines a weight function, thus $\hat{m}(\mathbb{P},\mathbb{Q}) = 0$ implies $\mathbb{P}\ R^\dagger\ \mathbb{Q}$. $\square$

By the way, the lifting operation given in Definition 3.1 can also be characterised as a maximum flow problem in a network [1].

4 Application in image retrieval

The *Earth Mover's distance* (EMD) was introduced by Rubner *et al.* [25,26] as an empirical way to measure color and texture similarities. It was shown to outperform many other texture similarity measures when used for texture classification and segmentation [23]. Formally, it is defined for "signatures" of the form $\{(x_1, p_1), ..., (x_m, p_m)\}$, where x_i is the center of data cluster i and p_i is the number of points in the cluster. Given two signatures $P = \{(x_1, p_1), ..., (x_n, p_n)\}$ and $Q = \{(y_1, q_1), ..., (y_m, q_m)\}$, whose total masses may not be equal, the EMD is defined in terms of the value of the linear programming problem:

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij} d(x_i, y_j) \\ \text{subject to} & \bullet\ \forall 1 \le i \le n : \sum_{j=1}^{m} f_{ij} \le p_i \\ & \bullet\ \forall 1 \le j \le m : \sum_{i=1}^{n} f_{ij} \le q_j \\ & \bullet\ \sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij} = \min(\sum_{i=1}^{n} p_i, \sum_{j=1}^{m} q_j) \\ & \bullet\ \forall 1 \le i \le n, 1 \le j \le m : f_{ij} \ge 0. \end{array} \tag{5}$$

where f_{ij} is the flow (the amount of earth moved) from cluster i to cluster j, and $d(x_i, y_j)$ is some measure of dissimilarity between x_i and y_j, say the Euclidean distance in $\mathcal{R}$. In the EMD terminology, the value of the objective function in (5) is the work required to move earth from one signature to another. Once the optimal flow f_{ij}^* is found, the Earth Mover's distance between P and Q is defined as

$$EMD(P, Q) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}^* d(x_i, y_j)}{\sum_{i=1}^{n}\sum_{j=1}^{m} f_{ij}^*} \tag{6}$$

For signatures with the same total mass the EMD is a true metric on distributions, and it is exactly the same as the Kantorovich metric, as noticed in [19] [2].

In [3], three other special forms of the Kantorovich metric were proposed to compare color histograms of images. Moreover, as a generalization of the problem of measuring image similarity, video clips can also be measured with the help of the Kantorovich metric, which turns out to be an effective technique [14].

5 Application in data mining

In [36] a Kantorovich distance based metric was proposed to compare clusterings. Given a dataset $D = \{x_1, ..., x_r\}$, suppose two clustering results Cls_1

[2] In [19] the Mallows metric is used, which is a special case of the Kantorovich metric.

and Cls_2 are obtained. They contain n and m clusters, respectively. Denote the n clusters in Cls_1 by $C_1, ..., C_n$, and the m clusters in Cls_2 by $C'_1, ..., C'_m$. Let the probability matrix generated by Cls_1 be $P = (p_{ij})$, where p_{ij} denotes the probability that object x_i belongs to cluster C_j. Let the corresponding matrix generated by Cls_2 be $Q = (q_{ij})$.

A cluster C_j is characterized by the r-dimensional vector $(p_{1j}, p_{2j}, ..., p_{rj})^T$, denoted by ξ_j. Similarly, denote the vector characterizing cluster C'_j by $\gamma_j = (q_{1j}, q_{2j}, ..., q_{r,j})^T$. To reflect the significance of each cluster for the purpose of comparison, we assign a weight to each cluster. Let the weight assigned to C_j be α_j with $\sum_{j=1}^{n} \alpha_j = 1$, and those to C'_j be β_j with $\sum_{j=1}^{m} \beta_j = 1$. So Cls_1 corresponds to the distribution $\mathbb{P} = \{(\xi_1, \alpha_1), ..., (\xi_n, \alpha_n)\}$ and Cls_2 to $\mathbb{Q} = \{(\gamma_1, \beta_1), ..., (\gamma_m, \beta_m)\}$. The distance between Cls_1 and Cls_2 is the value of the following linear programming problem:

$$\begin{array}{ll} \text{minimize} & \sum_{j=1}^{n} \sum_{k=1}^{m} w_{jk} \sum_{i=1}^{r} |p_{ij} - q_{ik}| \\ \text{subject to} & \bullet\ \forall 1 \leq j \leq n : \sum_{k=1}^{m} w_{jk} = \alpha_j \\ & \bullet\ \forall 1 \leq k \leq m : \sum_{j=1}^{n} w_{jk} = \beta_k \\ & \bullet\ \forall 1 \leq j \leq n, 1 \leq k \leq m : w_{jk} \geq 0. \end{array} \tag{7}$$

The clustering distance defined above is a Kantorovich distance.

6 Application in bioinformatics

An important problem in bioinformatics is to determine similarities and dissimilarities among DNA sequences, which can be used to detect the structural signature of a genome as well as to identify phylogenetic relationships among different species. One approach to solving this problem is based on statistical analysis of large DNA sequences using distribution of DNA words, which is a simple yet effective statistical tool to capture information about structural patterns, and these can reveal biologically significant features in a DNA sequence (see e.g. [2]). A DNA sequence is formed using an alphabet of four letters $\{A, T, C, G\}$ denoting four DNA bases: adenine, thymine, cytosine and guanine, respectively. A statistical summarization relies on various frequencies of DNA k-words, which are k-tuples formed via these four letters. Let $k \geq 1$ and W_k denote the set of all possible k-words formed using the alphabet $\{A, T, C, G\}$. Clearly, the size of the set W_k is 4^k. For a given DNA sequence and a word $w \in W_k$, let f_w be the relative frequency of the word w in the sequence, where the words in the sequence may have one or more overlapping letters. For example, in a sequence like $ATTCGGCA...$, the first 4-word is $ATTC$, the second one is $TTCG$, the third one is $TCGG$ and so on. The 4^k-dimensional frequency vector $(f_w)_{w \in W_k}$ constitutes a statistical summary of the given DNA sequence, and we have $\sum_{w \in W_k} f_w = 1$ with $f_w \geq 0$ for all $w \in W_k$. A comparison between a pair of DNA sequences to judge their similarities and dissimilarities can be carried out by comparing their associated frequency vectors, say $(f_w)_{w \in W_k}$ and $(g_w)_{w \in W_k}$. In [20] an appropriate metric for calculating the distance of two k-words is chosen. It is then lifted to be a metric over frequency vectors in terms of the Kantorovich metric. This idea has been implemented in a tool and some encouraging experimental results have been obtained.

7 Concluding remarks

We have briefly surveyed some recent applications of the Kantorovich metric in computer science. In general, production planning problems spread over a large variety of research areas, but lead to the same mathematical problem, namely the transportation problem. Therefore, the Kantorovich metric will probably find many more applications in the future. For example, this can be envisaged in relevant areas such as knowledge representation, statistical clustering, data mining, information retrieval, bioinformatics etc., all of which demand appropriate ways of computing similarity/dissimilarity between objects.

Acknowledgement

We thank the anonymous referees for insightful comments on an earlier version of the paper. Deng would like to acknowledge the support of the National Natural Science Foundation of China (Grant No. 60703033).

References

[1] C. Baier, B. Engelen, and M. E. Majster-Cederbaum. Deciding bisimilarity and similarity for probabilistic processes. *Journal of Computer and System Sciences*, 60(1):187–231, 2000.

[2] P. Chaudhuri and S. Das. Statistical analysis of large DNA sequences using distribution of DNA words. *Current Science*, 80(9):1161–1166, 2001.

[3] K. Chen, S. Demko, and R. Xie. Similarity-based retrieval of images using color histograms. In *Storage and Retrieval for Image and Video Databases VII*, volume 3656 of *SPIE*, pages 643–652, 1999.

[4] Y. Deng, T. Chothia, C. Palamidessi, and J. Pang. Metrics for action-labelled quantitative transition systems. *Electronic Notes in Theoretical Computer Science*, 153(2):79–96, 2006.

[5] Y. Deng and W. Du. Probabilistic barbed congruence. *Electronic Notes in Theoretical Computer Science*, 190(3):185–203, 2007.

[6] Y. Deng, R. van Glabbeek, M. Hennessy, and C. Morgan. Characterising testing preorders for finite probabilistic processes. *Logical Methods in Computer Science*, 4(4:4):1–33, 2008.

[7] Y. Deng, R. van Glabbeek, M. Hennessy, C. Morgan, and C. Zhang. Characterising testing preorders for finite probabilistic processes. In *Proceedings of the 22nd Annual IEEE Symposium on Logic in Computer Science*, pages 313–325. IEEE Computer Society, 2007.

[8] J. Desharnais, R. Jagadeesan, V. Gupta, and P. Panangaden. Metrics for labeled Markov systems. In *Proceedings of the 10th International Conference on Concurrency Theory*, volume 1664 of *Lecture Notes in Computer Science*, pages 258–273. Springer-Verlag, 1999.

[9] J. Desharnais, R. Jagadeesan, V. Gupta, and P. Panangaden. The metric analogue of weak bisimulation for probabilistic processes. In *Proceedings of the 17th Annual IEEE Symposium on Logic in Computer Science*, pages 413–422. IEEE Computer Society, 2002.

[10] J. Desharnais, R. Jagadeesan, V. Gupta, and P. Panangaden. Metrics for labelled Markov processes. *Theoretical Computer Science*, 318(3):323–354, 2004.

[11] N. Ferns, P. Panangaden, and D. Precup. Metrics for finite Markov decision processes. In *Proceedings of the 20th Conference in Uncertainty in Artificial Intelligence*, pages 162–169. AUAI Press, 2004.

[12] N. Ferns, P. Panangaden, and D. Precup. Metrics for Markov decision processes with infinite state spaces. In *Proceedings of the 21st Conference in Uncertainty in Artificial Intelligence*, pages 201–208. AUAI Press, 2005.

[13] A. L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, 2002.

[14] C. Grana, D. Borghesani, and R. Cucchiara. Video shots comparison using the Mallows distance. In *Proceedings of the 18th International Conference on Database and Expert Systems Applications*, pages 49–53. IEEE Computer Society, 2007.

[15] B. Jonsson and K. Larsen. Specification and refinement of probabilistic processes. In *Proceedings of the 6th Annual IEEE Symposium on Logic in Computer Science*, pages 266–277. Computer Society Press, 1991.

[16] L. V. Kantorovich. On the translocation of masses. *Doklady Akademii Nauk SSSR*, 37(7-8):227–229, 1942.

[17] L. V. Kantorovich and G. S. Rubinshtein. On a space of totally additive functions. *Vestn Lening. Univ.*, 13(7):52–59, 1958.

[18] K. G. Larsen and A. Skou. Bisimulation through probabilistic testing. *Information and Computation*, 94(1):1–28, 1991.

[19] E. Levina and P. J. Bickel. The Earth Mover's distance is the Mallows distance: Some insights from statistics. In *Proceedings of the 8th International Conference On Computer Vision*, volume 2, pages 251–256. IEEE Computer Society, 2001.

[20] X. Li and Y. Deng. Automatic measurement of large DNA sequences, 2009. In preparation.

[21] G. Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Academie des Science de Paris*, page 666, 1781.

[22] J. B. Orlin. A faster strongly polynomial minimum cost flow algorithm. In *Proceedings of the 20th ACM Symposium on the Theory of Computing*, pages 377–387. ACM, 1988.

[23] J. Puzicha, Y. Rubner, C. Tomasi, and J. Buhmann. Empirical evaluation of dissimilarity measures for color and texture. In *Proceedings of the 6th International Conference On Computer Vision*, pages 1165–1173. IEEE Computer Society, 1999.

[24] S. Rachev. *Probability Metrics and the Stability of Stochastic Models*. Wiley New York, 1991.

[25] Y. Rubner, C. Tomasi, and L. Guibas. The Earth Mover's distance as a metric for image retrieval. Technical Report STAN-CS-TN-98-86, Department of Computer Science, Stanford University, 1998.

[26] Y. Rubner, C. Tomasi, and L. Guibas. A metric for distributions with applications to image databases. In *Proceedings of the 5th International Conference on Computer Vision*, pages 59–66. IEEE Computer Society, 1998.

[27] F. van Breugel, C. Hermida, M. Makkai, and J. Worrell. An accessible approach to behavioural pseudometrics. In *Proceedings of the 32nd International Colloquium on Automata, Languages and Programming*, volume 3580 of *Lecture Notes in Computer Science*, pages 1018–1030. Springer, 2005.

[28] F. van Breugel, C. Hermida, M. Makkai, and J. Worrell. Recursively defined metric spaces without contraction. *Theoretical Computer Science*, 380(1-2):143–163, 2007.

[29] F. van Breugel, B. Sharma, and J. Worrell. Approximating a behavioural pseudometric without discount for probabilistic systems. In *Proceedings of the 10th International Conference on Foundations of Software Science and Computational Structures*, volume 4423 of *Lecture Notes in Computer Science*, pages 123–137. Springer, 2007.

[30] F. van Breugel and J. Worrell. An algorithm for quantitative verification of probabilistic transition systems. In *Proceedings of the 12th International Conference on Concurrency Theory*, volume 2154 of *Lecture Notes in Computer Science*, pages 336–350. Springer-Verlag, 2001.

[31] F. van Breugel and J. Worrell. Towards quantitative verification of probabilistic transition systems. In *Proceedings of the 28th International Colloquium*

Automata, Languages and Programming, volume 2076 of *Lecture Notes in Computer Science*, pages 421–432. Springer-Verlag, 2001.

[32] F. van Breugel and J. Worrell. A behavioural pseudometric for probabilistic transition systems. *Theoretical Computer Science*, 331(1):115–142, 2005.

[33] F. van Breugel and J. Worrell. Approximating and computing behavioural distances in probabilistic transition systems. *Theoretical Computer Science*, 360(1-3):373–385, 2006.

[34] A. Vershik. Kantorovich metric: Initial history and little-known applications. *Journal of Mathematical Sciences*, 133(4):1410–1417, 2006.

[35] C. Villani. *Topics in Optimal Transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, 2003.

[36] D. Zhou, J. Li, and H. Zha. A new Mallows distance based metric for comparing clusterings. In *Proceedings of the Twenty-Second International Conference on Machine Learning*, volume 119, pages 1028–1035. ACM, 2005.

[37] K. Życzkowski and W. Slomczyński. The Monge distance between quantum states. *Journal of Physics A: Mathematical and General*, 31(45):9095–9104, 1998.

A Mathematical properties of the pseudometric $\hat{m}$

Given the ground distance function m, the pseudometric $\hat{m}$ on distributions (cf. Section 3.1) enjoys some simple but useful mathematical properties which we list below.

(i) (Nonnegativity) $\hat{m}(\mathbb{P},\mathbb{Q}) \geq 0$, with $\hat{m}(\mathbb{P},\mathbb{Q}) = 0$ if $\mathbb{P} = \mathbb{Q}$.

(ii) (Symmetry) $\hat{m}(\mathbb{P},\mathbb{Q}) = \hat{m}(\mathbb{Q},\mathbb{P})$.

(iii) (Triangle inequality) $\hat{m}(\mathbb{P},\mathbb{R}) \leq \hat{m}(\mathbb{P},\mathbb{Q}) + \hat{m}(\mathbb{Q},\mathbb{R})$.

(iv) (Possibility of extension) $\hat{m}(\mathbb{P},\mathbb{Q}) = \hat{m}(\mathbb{P}',\mathbb{Q}')$ where $dom(\mathbb{P}') = dom(\mathbb{P}) \cup \{s\}$ and $\mathbb{P}'(s) = 0$, similarly for $\mathbb{Q}'$ with respect to $\mathbb{Q}$.

(v) (Convexity) For $0 \leq p \leq 1$, we have

$$\hat{m}(p \cdot \mathbb{P} + (1-p) \cdot \mathbb{R},\ p \cdot \mathbb{Q} + (1-p) \cdot \mathbb{R}) = p \cdot \hat{m}(\mathbb{P},\mathbb{Q}).$$

(vi) (Joint convexity) For $0 \leq p \leq 1$, we have

$$\hat{m}(p \cdot \mathbb{P} + (1-p) \cdot \mathbb{Q},\ p \cdot \mathbb{P}' + (1-p) \cdot \mathbb{Q}') \ \leq\ p \cdot \hat{m}(\mathbb{P},\mathbb{P}') + (1-p) \cdot \hat{m}(\mathbb{Q},\mathbb{Q}').$$