A hierarchy of convex relaxations for the total variation distance

Jean-Bernard Lasserre

A HIERARCHY OF CONVEX RELAXATIONS FOR THE TOTAL VARIATION DISTANCE

JEAN B. LASSERRE

Abstract. Given two measures μ, ν on $\mathbb{R}^d$ that satisfy Carleman's condition, we provide a numerical scheme to approximate as closely as desired the total variation distance between μ and ν. It consists of solving a sequence (hierarchy) of convex relaxations whose associated sequence of optimal values converges to the total variation distance, an additional illustration of the versatility of the Moment-SOS hierarchy. Indeed each relaxation in the hierarchy is a semidefinite program whose size increases with the number of involved moments. It has an optimal solution which is a couple of degree-$2n$ pseudo-moments which converge, as n grows, to moments of the Hahn-Jordan decomposition of $\mu - \nu$.

1. Introduction

Evaluating a "distance" between measures has become an important topic with many applications, especially in Data Science and Machine Learning in particular. Among possible choices, the Wasserstein distance has become popular and one reason is that its optimal transport formulation allows to define efficient specialized procedures (e.g. the Sinkhorn algorithm) for its computation [10].

Contribution. In this paper we show that the *total variation* distance is also amenable to practical computation under fairly weak assumptions and so could provide an alternative to other distances when needed; see e.g. [9] for a discussion on relative merits of several distances. The total variation distance being the same as the Wasserstein distance with (nasty) cost function $c(x, y) = 1_{x \neq y}(x, y)$, is an indication that its effective computation is a computational challenge.

An important application of the total variation is in computer vision where it was introduced in [11] as a regularization criterion in some inverse problems (e.g. denoising of images), followed by [3]. Next, [2] inspired by [3], has proposed an algorithm to minimize the (discrete) total variation of functions $u \in L^1(\Omega)$ for a compact set $\Omega \subset \mathbb{R}^2$ (which in principle can be adapted to higher dimensions). (A discretized image is seen a $N \times N$ matrix.)

Our focus is different from the above cited works [2, 3, 11]. Independently of a particular application, and in a rather general context, we provide a numerical scheme to approximate as closely as desired the total variation distance between two measures μ and ν. We assume that all moments of μ and ν are finite, and that both

The author is supported by the AI Interdisciplinary Institute ANITI funding through the french program "Investing for the Future PI3A" under the grant agreement number ANR-19-PI3A-0004. This research is also part of the programme DesCartes and is supported by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme.

μ and ν satisfy Carleman' condition. In particular we do *not* assume that μ or ν has compact support. We can formulate the problem as an instance of the Generalized Moment Problem and show that it is amenable to practical computation via the Moment-SOS hierarchy [7, 4]. As a result, one may approximate as closely as desired $\|\mu-\nu\|_{TV}$ as more and more moments of μ and ν are taken into account. Our contribution is to provide an additional tool in the arsenal of algorithms available in applied probability, for approximating as closely as desired, the total variation distance $\|\mu-\nu\|_{TV}$ based on moment information.

(i) This numerical scheme consists of solving a sequence (hierarchy) of convex relaxations. Each convex relaxation of the hierarchy is a semidefinite program[1] whose size increases with the number of moments of μ and ν involved.

(ii) The associated sequence of optimal values is monotone non decreasing and converges from below to $\|\mu-\nu\|_{TV}$.

(iii) At last but not least, the associated sequence of optimal solutions of relaxations (a couple of vectors whose size increases), converges to the couple of infinite moment-vectors of the Hahn-Jordan decomposition (ϕ^*_+,ϕ^*_-) of the signed measure $\mu-\nu$.

We wish to emphasize the weak assumption on the measures μ,ν, namely that they satisfy Carleman's condition (no compact support is required). It is a feature of the total variation distance to not discriminate between mutually singular probability measures (their distance is constant to 1). Relatively surprisingly, and as a good sign of the numerical scheme's behavior, it is shown in Example 1 that the exact distance $\|\delta_0-\delta_\varepsilon\|_{TV}$ between the two (mutually singular) Dirac measures at 0 and $\varepsilon>0$, is obtained at the first semidefinite relaxation of the hierarchy, irrespective of the value of $\varepsilon>0$ (whereas one might have expected that convergence of the relaxations would depend on ε).

Interestingly, and as an alternative to algorithms based on a discretization (like e.g. Sinkhorn algorithm), the Wasserstein distance $W_2(\mu,\nu)$ (with *polynomial* cost $c(x,y)$) can also be approximated as closely as desired by a mesh-free practical computation by (i) applying the Moment-SOS hierarchy [4, 8] for solving the associated optimal transport problem (OT), and (ii) extract the transport map from the moment vector solution of the OT, by a non-standard application of the Christoffel-Darboux kernel [5]. However, crucial is the fact that the cost function is a polynomial (which of course excludes the nasty cost function $1_{x\neq y}(x,y)$).

2. Main result

2.1. **Notation and definitions.** Let $\mathbb{R}[\mathbf{x}]$ denote the ring of real polynomials in the variables $(x_1,\ldots,x_d)$ and $\mathbb{R}[\mathbf{x}]_n\subset\mathbb{R}[\mathbf{x}]$ be its subset of polynomials of total degree at most n. Let $\mathbb{N}^d_n:=\{\boldsymbol{\alpha}\in\mathbb{N}^d:\sum_i\alpha_i\leq n\}$ with cardinal $s(n)=\binom{n+d}{n}$. Let $\mathbf{v}_n(\mathbf{x})=(x^{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^d_n}$ be the vector of monomials up to degree n, and let $\Sigma[\mathbf{x}]_n\subset\mathbb{R}[\mathbf{x}]_{2n}$ be the convex cone of polynomials of total degree at most $2n$ which are sum-of-squares (in short SOS). A polynomial $p\in\mathbb{R}[\mathbf{x}]_n$ can be identified with its vector of coefficients $\mathbf{p}=(p_{\boldsymbol{\alpha}})\in\mathbb{R}^{s(n)}$ in the monomial basis, and reads

$$\mathbf{x}\quad\mapsto p(\mathbf{x}):=\langle\mathbf{p},\mathbf{v}_n(\mathbf{x})\rangle\,,\quad\forall p\in\mathbb{R}[\mathbf{x}]\,.$$

[1] A semidefinite program is a convex conic optimization problem that can be solved efficiently, up to arbitrary precision fixed in advance; see e.g. [1]

Denote by $\mathscr{M}(\mathbb{R}^d)$ (resp. $\mathscr{M}(\mathbb{R}^d)_+$) the space of signed (resp. positive) Borel measures on $\mathbb{R}^d$. For two Borel measures $\mu, \nu \in \mathscr{M}(\mathbb{R}^d)_+$, the notation $\mu \leq \nu$ stands for $\mu(B) \leq \nu(B)$ for all Borel sets $B \in \mathcal{B}(\mathbb{R}^d)$. The support of a Borel measure μ on $\mathbb{R}^d$ is the smallest closed set A such that $\mu(\mathbb{R}^d \setminus A) = 0$, and such a set A is unique. A Borel measure whose all moments are finite is said to be (moment) *determinate* if there is no other measure with same moments.

For a real symmetric matrix $\mathbf{A} = \mathbf{A}^T$, the notation $\mathbf{A} \succeq 0$ (resp. $\mathbf{A} \succ 0$) stands for $\mathbf{A}$ is positive semidefinite (p.s.d.) (resp. positive definite (p.d.)).

Hahn-Jordan decomposition. Given two finite Borel measures $\mu, \nu \in \mathscr{M}(\mathbb{R}^d)_+$, the signed measure $\mu - \nu$ has a unique Hahn-Jordan decomposition (ϕ^*_+, ϕ^*_-) such that $\phi^*_+ - \phi^*_- = \mu - \nu$. That is, there exists a Borel set $A \in \mathcal{B}(\mathbb{R}^d)$ and two mutually singular positive measures ϕ^*_+, ϕ^*_- such that $\phi^*_+(\mathbb{R}^d) = \phi^*_+(A)$ while $\phi^*_-(A) = 0$, and

$$(2.1)\quad \phi^*_+(B) \,=\, (\mu-\nu)(B\cap A)\,; \quad \phi^*_-(B) \,=\, (\nu-\mu)(B\cap(\mathbb{R}^d\setminus A))\,, \quad \forall B \in \mathcal{B}(\mathbb{R}^d)\,.$$

In addition, and obviously, $\|\mu-\nu\|_{TV} \leq \mu(1)+\nu(1)$. Moreover, observe that $\phi^*_+ \leq \mu$ and $\phi^*_- \leq \nu$.

Riesz linear functional and moment matrix. With a real sequence $\boldsymbol{\phi} = (\phi_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^d}$ (in bold) is associated the *Riesz* linear functional $\phi \in \mathbb{R}[\mathbf{x}]^*$ (not in bold) defined by

$$p\,(=\sum_{\boldsymbol{\alpha}} p_{\boldsymbol{\alpha}}\mathbf{x}^{\boldsymbol{\alpha}}) \quad \mapsto \phi(p) \,=\, \langle \boldsymbol{\phi}, \mathbf{p}\rangle \,=\, \sum_{\boldsymbol{\alpha}} p_{\boldsymbol{\alpha}}\,\phi_{\boldsymbol{\alpha}}\,, \quad \forall p \in \mathbb{R}[\mathbf{x}]\,,$$

and the moment matrix $\mathbf{M}_n(\boldsymbol{\phi})$ with rows and columns indexed by $\mathbb{N}^d_n$ (hence of size $s(n)$), and with entries

$$\mathbf{M}_n(\boldsymbol{\phi})(\boldsymbol{\alpha},\boldsymbol{\beta}) \,:=\, \phi(\mathbf{x}^{\boldsymbol{\alpha}+\boldsymbol{\beta}}) \,=\, \phi_{\boldsymbol{\alpha}+\boldsymbol{\beta}}\,, \quad \boldsymbol{\alpha},\boldsymbol{\beta} \in \mathbb{N}^d_n\,.$$

Notice that one may write indifferently $\mathbf{M}_n(\boldsymbol{\phi})$ or $\mathbf{M}_n(\phi)$, i.e., referring to the sequence $\boldsymbol{\phi}$ truncated to degree-$2n$ moments or to the Riesz linear functional ϕ associated with $\boldsymbol{\phi}$.

A real sequence $\boldsymbol{\phi} = (\phi_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^d}$ has a representing mesure if its associated linear functional ϕ is a Borel measure on $\mathbb{R}^d$. In this case $\mathbf{M}_n(\boldsymbol{\phi}) \succeq 0$ for all n; the converse is not true in general.

Carleman's condition. A sequence $\boldsymbol{\mu} = (\mu_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^d}$ satisfies Carleman's condition if

$$(2.2)\qquad \forall i = 1,\ldots,d\,: \quad \sum_{j=1}^{\infty} \mu(x_i^{2j})^{-1/2j} \,=\, +\infty\,.$$

The following theorem is due to Nussbaum:

Theorem 2.1. *(*[8, Theorem 3.5]*) Let a sequence $\boldsymbol{\mu} = (\mu_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^d}$ be such that $\mathbf{M}_n(\boldsymbol{\mu}) \succeq 0$, for all $n \in \mathbb{N}$. If $\boldsymbol{\mu}$ satisfies Carleman's condition* (2.2) *then $\boldsymbol{\mu}$ has a representing measure μ on $\mathbb{R}^d$ and μ is determinate.*

A sufficient condition to ensure that a measure μ satisfies the multivariate Carleman's condition is that

$$\int \exp(c|x_i|)\,d\mu \,<\, \infty\,, \quad i=1,\ldots,d\,, \tag{2.3}$$

if for some scalar $c>0$.

2.2. **A preliminary result.**

Lemma 2.2. *Let $\mu,\varphi\in\mathscr{M}(\mathbb{R}^d)_+$ have finite moments and assume that μ satisfies Carleman's condition* (2.2)*. Then*

$$\varphi\le\mu \quad\Leftrightarrow\quad \mathbf{M}_n(\varphi)\preceq\mathbf{M}_n(\mu)\,,\quad \forall n\in\mathbb{N}\,. \tag{2.4}$$

Proof. $\Rightarrow$ is straightforward. Indeed:

$$\mu\ge\varphi\Rightarrow\left[\int p^2\,d\mu\,\ge\,\int p^2\,d\varphi\,,\,\forall p\in\mathbb{R}[\mathbf{x}]\right]\quad\Rightarrow\quad \mathbf{M}_n(\mu)\,\succeq\,\mathbf{M}_n(\varphi)\,,\,\forall n\in\mathbb{N}\,.$$

$\Leftarrow$ Assume that $\mathbf{M}_n(\varphi)\preceq\mathbf{M}_n(\mu)$ for all $n\in\mathbb{N}$, and consider the sequence $\boldsymbol{\gamma}=(\gamma_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^d}$, with $\gamma_{\boldsymbol{\alpha}}=\mu_{\boldsymbol{\alpha}}-\varphi_{\boldsymbol{\alpha}}$, for all $\boldsymbol{\alpha}\in\mathbb{N}^d$. Then $\int x_i^{2n}d\varphi\le\int x_i^{2n}d\mu$ for all n, and as Carleman's condition (2.2) holds for μ, we infer $\gamma(x_i^{2n})\le\mu(x_i^{2n})$ for all n, and all $i=1\ldots,d$. This implies that $\boldsymbol{\gamma}$ satisfies Carleman's condition (2.2) and therefore, as $\mathbf{M}_n(\boldsymbol{\gamma})=\mathbf{M}_n(\mu)-\mathbf{M}_n(\varphi)\succeq 0$ for all n, we deduce that $\boldsymbol{\gamma}$ has a determinate representing measure γ on $\mathbb{R}^d$. In particular:

$$\int\mathbf{x}^{\boldsymbol{\alpha}}\,d(\gamma+\varphi)\,=\,\gamma_{\boldsymbol{\alpha}}+\varphi_{\boldsymbol{\alpha}}\,=\,\mu_{\boldsymbol{\alpha}}\,=\,\int\mathbf{x}^{\boldsymbol{\alpha}}\,d\mu\,,\quad\forall\boldsymbol{\alpha}\in\mathbb{N}^d\quad\Rightarrow\gamma+\varphi\,=\,\mu\,,$$

where the last statement follows from determinateness of μ. Hence $\varphi\le\mu$. □

2.3. **Main result.** Given two finite Borel measures μ and ν on $\mathbb{R}^d$, introduce the infinite-dimensional LP:

$$\tau=\inf_{\phi^+,\phi^-\in\mathscr{M}(\mathbb{R}^d)_+}\{\,\phi^+(1)+\phi^-(1):\quad\phi_+-\phi_-\,=\,\mu-\nu\,\}\,. \tag{2.5}$$

Proposition 2.3. *The LP* (2.5) *has a unique optimal solution (ϕ_+^*,ϕ_-^*) which is the Hahn-Jordan decomposition of the signed measure $\mu-\nu$, and therefore $\tau=\phi_+^*(1)+\phi_-^*(1)=\|\mu-\nu\|_{TV}$.*

Proof. Let (ϕ^+,ϕ^-) be an arbitrary feasible solution of (2.5). Then as $\phi^+-\phi^-=\mu-\nu$ one obtains $\phi^+(1)+\phi^-(1)\ge\|\phi^+-\phi^-\|_{TV}=\|\mu-\nu\|_{TV}$. On the other hand, the Hahn-Jordan decomposition (ϕ_+^*,ϕ_-^*) of $\mu-\nu$ is feasible for (2.5), with value $\|\mu-\nu\|_{TV}$, whence the result. □

Unfortunately the LP (2.5) is not very useful as its stands. It is just a particular rephrasing of the total variation distance between μ and ν. However we next see the a slight reinforcement of (2.5) will turn out to be very useful when passing to some hierarchy of convex relaxations. Indeed:

Proposition 2.4. *The linear program*

$$\inf_{\phi^+,\phi^-\in\mathscr{M}(\mathbb{R}^d)_+}\{\,\phi^+(1)+\phi^-(1):\quad\phi_+-\phi_-\,=\,\mu-\nu\,;\quad\phi^+\le\mu\,;\,\phi^-\le\nu\,\} \tag{2.6}$$

has same optimal value $\tau=\|\mu-\nu\|_{TV}$, and optimal solution (ϕ_+^*,ϕ_-^*) as (2.5)*.*

Proof. By construction, the optimal value ρ of (2.6) satisfies $\rho \geq \tau = \|\mu - \nu\|_{TV}$. On the other hand, with (ϕ_+^*, ϕ_-^*) being the Hahn-Jordan decomposition of $\mu - \nu$, observe that $\phi_+^* \leq \mu$, and $\phi_-^* \leq \nu$. Therefore (ϕ_+^*, ϕ_-^*) is an optimal solution of (2.6). Equivalently, the constraints $\phi^+ \leq \mu$ and $\phi^- \leq \nu$ are automatically satisfied at the optimal solution (ϕ_+^*, ϕ_-^*) of (2.5) and therefore (2.5) and (2.6) have same optimal value and same optimal solution. □

Next, from now on we make the following assumption:

Assumption 2.5. *(i) All moments of μ and ν are finite, and*

(ii) μ and ν satisfy (2.3) *(hence satisfy Carleman's condition* (2.2)*) for some scalar $c > 0$.*

Consider the optimization problem

(2.7)
$$\hat{\tau} = \min_{\phi^+, \phi^- \in \mathscr{M}(\mathbb{R}^d)_+} \{\, \phi^+(1) + \phi^-(1) : \quad \phi^+ - \phi^- = \mu - \nu\,; \quad \mathbf{M}_n(\phi^+) \preceq \mathbf{M}_n(\mu)\,; \quad \mathbf{M}_n(\phi^-) \preceq \mathbf{M}_n(\nu)\,, \quad \forall n \in \mathbb{N}\,\}.$$

Corollary 2.6. *Let Assumption 2.5 hold. Then the Hahn-Jordan decomposition (ϕ_+^*, ϕ_-^*) of the signed measure $\mu - \nu$, is the unique optimal solution of* (2.7)*, and $\hat{\tau} = \tau = \|\mu - \nu\|_{TV}$.*

Proof. By Lemma 2.2, (2.6) and (2.7) are equivalent. □

The nice feature of the LP (2.7) when compared to its equivalent formulation (2.6), is that the cost as well as the constraints of (2.7) can next be formulated in terms of moments of $(\mu, \nu, \phi^+, \phi^-)$, so as to yield the optimization problem:

$$\rho = \min_{\phi^+, \phi^- \in \mathscr{M}(\mathbb{R}^d)_+} \{\, \phi^+(1) + \phi^-(1) : \int \mathbf{x}^{\boldsymbol{\alpha}} d(\phi^+ - \phi^-) = \int \mathbf{x}^{\boldsymbol{\alpha}}\, d(\mu - \nu)\,, \quad \forall \boldsymbol{\alpha} \in \mathbb{N}^d \quad \mathbf{M}_n(\phi^+) \preceq \mathbf{M}_n(\mu)\,;\, \mathbf{M}_n(\phi^-) \preceq \mathbf{M}_n(\nu)\,,\, \forall n \in \mathbb{N}\,\}\,, \tag{2.8}$$

which is an instance of the Generalized Moment Problem (GMP); see e.g. [8].

Corollary 2.7. *Let Assumption 2.5 hold. Then the Hahn-Jordan decomposition (ϕ_+^*, ϕ_-^*) of the signed measure $\mu - \nu$, is the unique optimal solution of* (2.8)*, and $\rho = \|\mu - \nu\|_{TV}$.*

Proof. Let (ϕ^+, ϕ^-) be an arbitrary feasible solution of (2.8). By Lemma 2.2, $\phi^+ \leq \mu$ and $\phi^- \leq \nu$. Hence $\phi^+ + \nu \leq \mu + \nu$, and $\phi^- + \mu \leq \mu + \nu$. As Assumption 2.5(ii) holds,

$$\begin{aligned} \int \exp(c\,|x_i|)\, d(\phi^+ + \nu) &< \int \exp(c|x_i|)\, d(\mu + \nu) < \infty \\ \int \exp(c\,|x_i|)\, d(\phi^- + \mu) &< \int \exp(c|x_i|)\, d(\mu + \nu) < \infty\,, \end{aligned}$$

and therefore the measure $\phi^+ + \nu$ (resp. $\phi^- + \mu$) is determinate. But then the constraint $\int \mathbf{x}^{\boldsymbol{\alpha}} d(\phi^+ - \phi^-) = \int \mathbf{x}^{\boldsymbol{\alpha}} d(\mu - \nu)$ for all $\boldsymbol{\alpha} \in \mathbb{N}^d$ reads:

$$\int \mathbf{x}^{\boldsymbol{\alpha}}\, d(\phi^+ + \nu) = \int \mathbf{x}^{\boldsymbol{\alpha}}\, d(\phi^- + \mu)\,, \quad \forall \boldsymbol{\alpha} \in \mathbb{N}^d\,,$$

which implies $\phi^+ + \nu = \phi^- + \mu$ by determinacy of the measures. Therefore (ϕ^+, ϕ^-) is a feasible solution of (2.7) with same value. In other words, (2.8) is equivalent to (2.7), whence the result. □

2.4. **A convergent hierarchy of semidefinite relaxations.** As (2.8) is an instance of the GMP, it is natural to apply the Moment-SOS hierarchy [4, 7]. With each fixed $n \in \mathbb{N}$, consider the optimization problem

$$\rho_n = \min_{\boldsymbol{\phi},\boldsymbol{\psi}} \quad \{\, \phi(1) + \psi(1) : \quad \phi_{\boldsymbol{\alpha}} - \psi_{\boldsymbol{\alpha}} \,=\, \mu_{\boldsymbol{\alpha}} - \nu_{\boldsymbol{\alpha}}\,, \quad \forall \boldsymbol{\alpha} \in \mathbb{N}^d_{2n}\,; \\ 0 \preceq \mathbf{M}_n(\boldsymbol{\phi}) \preceq \mathbf{M}_n(\boldsymbol{\mu})\,; \quad 0 \preceq \mathbf{M}_n(\boldsymbol{\psi}) \preceq \mathbf{M}_n(\boldsymbol{\nu})\,\}\,, \tag{2.9}$$

where now the optimization is over degree-$2n$ pseudo-moment vectors $\boldsymbol{\phi} = (\phi_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^d_{2n}}$ and $\boldsymbol{\psi} = (\psi_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^d_{2n}}$ (hence not necessarily coming from measures ϕ and ψ on $\mathbb{R}^d$). Of course (2.9) is an obvious relaxation of (2.8) and therefore $\rho_n \le \rho = \|\mu - \nu\|_{TV}$ for all $n \in \mathbb{N}$.

Observe that for each fixed $n \in \mathbb{N}$, (2.9) is a semidefinite program that can be solved by off-the-shelf solvers like GloptiPoly [6] or Jump [12] (package of the Julia programming language).

Theorem 2.8. *Let Assumption 2.5 hold.*

(i) For every fixed $n \in \mathbb{N}$, the optimization problem (2.9) *has an optimal solution denoted $(\boldsymbol{\phi}^{(n)}, \boldsymbol{\psi}^{(n)})$.*

(ii) In addition, $\rho_n \uparrow \|\mu - \nu\|_{TV}$ as $n \to \infty$, and moreover,

$$\lim_{n\to\infty} \phi^{(n)}_{\boldsymbol{\alpha}} \,=\, \int \mathbf{x}^{\boldsymbol{\alpha}}\, d\phi^*_+\,; \quad \lim_{n\to\infty} \psi^{(n)}_{\boldsymbol{\alpha}} \,=\, \int \mathbf{x}^{\boldsymbol{\alpha}}\, d\phi^*_-\,, \quad \forall \boldsymbol{\alpha} \in \mathbb{N}^d\,, \tag{2.10}$$

where (ϕ^*_+, ϕ^*_-) is the Hahn-Jordan decomposition of the signed measure $\mu - \nu$.

Proof. (i) Let $(\boldsymbol{\phi}, \boldsymbol{\psi})$ be an arbitrary feasible solution of (2.9). As $\mathbf{M}_n(\boldsymbol{\phi}) \preceq \mathbf{M}_n(\mu)$ one obtains

$$\phi(1) \,\le\, \mu(1)\,; \quad \phi(x_i^{2n}) \,\le\, \mu(x_i^{2n})\,, \quad \forall i = 1, \ldots, d\,,$$

and therefore, as $\mathbf{M}_n(\phi^+) \succeq 0$, by [8, Proposition 3.6],

$$|\phi_{\boldsymbol{\alpha}}| \,\le\, \max[\mu(1)\,, \max_i \mu(x_i^{2d})]\,, \quad \forall \boldsymbol{\alpha} \in \mathbb{N}^d_{2n}\,. \tag{2.11}$$

Similarly, as $0 \preceq \mathbf{M}_n(\boldsymbol{\psi}) \preceq \mathbf{M}_n(\nu)$,

$$|\psi_{\boldsymbol{\alpha}}| \,\le\, \max[\nu(1)\,, \max_i \nu(x_i^{2d})]\,, \quad \forall \boldsymbol{\alpha} \in \mathbb{N}^d_{2n}\,. \tag{2.12}$$

Therefore the feasible set of (2.9) is compact. Hence (2.9) has an optimal solution.

(ii) For each fixed $n \in \mathbb{N}$, and since $\mathbf{M}_k(\boldsymbol{\phi}^{(n)})$ is a submatrix of $\mathbf{M}_n(\boldsymbol{\phi}^{(n)})$ for all $k = 1, \ldots, n$, again by [8,],

$$\forall \boldsymbol{\alpha} : 2k-1 \le |\boldsymbol{\alpha}| \le 2k : \quad |\phi^{(n)}_{\boldsymbol{\alpha}}| \,\le\, \max[\mu(1)\,, \max_i \mu(x_i^{2k})] \,=:\, a_k,; \quad k = 1, \ldots, n\,,$$

and similarly

$$|\psi^{(n)}_{\boldsymbol{\alpha}}| \,\le\, \max[\nu(1)\,, \max_i \nu(x_i^{2k})] \,=:\, b_k, \quad \forall \boldsymbol{\alpha} \,:\, 2k-1 \le |\boldsymbol{\alpha}| \le 2k\,;\, k = 1, \ldots, n\,.$$

Next, introduce the new infinite peudo-moment sequences:

$$\hat{\phi}^{(n)}_{\boldsymbol{\alpha}} \,:=\, \phi^{(n)}_{\boldsymbol{\alpha}}/a_k\,, \quad \forall \boldsymbol{\alpha} :\, 2k-1 \le |\boldsymbol{\alpha}| \le 2k\,; \quad k = 1, \ldots, n\,, \tag{2.13}$$

and $\hat{\phi}^{(n)}_{\boldsymbol{\alpha}} = 0$ for all $\boldsymbol{\alpha} \in \mathbb{N}^d$ with $|\boldsymbol{\alpha}| > 2n$. Similarly,

$$\hat{\psi}^{(n)}_{\boldsymbol{\alpha}} \,:=\, \psi^{(n)}_{\boldsymbol{\alpha}}/b_k\,, \quad \forall \boldsymbol{\alpha} :\, 2k-1 \le |\boldsymbol{\alpha}| \le 2k\,; \quad k = 1, \ldots, n\,, \tag{2.14}$$

and $\hat{\psi}^{(n)}_{\boldsymbol{\alpha}} = 0$ for all $\boldsymbol{\alpha} \in \mathbb{N}^d$ with $|\boldsymbol{\alpha}| > 2n$.

Both sequences $\hat{\boldsymbol{\phi}}^{(n)}$ and $\hat{\boldsymbol{\psi}}^{(n)}$ are considered as elements of the unity ball $\mathbf{B}(0,1)$ of the Banach space ℓ_∞ of uniformly bounded sequences, which is sequentially compact in the $\sigma(\ell_\infty,\ell_1)$ weak topology. Therefore there exist $\hat{\boldsymbol{\phi}},\hat{\boldsymbol{\psi}}\in\mathbf{B}(0,1)$ and a subsequence $(n_k)_{k\in\mathbb{N}}$ such that

$$\lim_{k\to\infty}\hat{\phi}_{\boldsymbol{\alpha}}^{(n_k)} = \hat{\phi}_{\boldsymbol{\alpha}}\,; \quad \lim_{k\to\infty}\hat{\psi}_{\boldsymbol{\alpha}}^{(n_k)} = \hat{\psi}_{\boldsymbol{\alpha}}\,, \quad \forall\boldsymbol{\alpha}\in\mathbb{N}^d\,. \tag{2.15}$$

By doing the reverse scaling of (2.13)-(2.14), one obtains:

$$\forall\boldsymbol{\alpha}\in\mathbb{N}^d: \quad \lim_{k\to\infty}\phi_{\boldsymbol{\alpha}}^{(n_k)} = \phi_{\boldsymbol{\alpha}}\,; \quad \lim_{k\to\infty}\psi_{\boldsymbol{\alpha}}^{(n_k)} = \psi_{\boldsymbol{\alpha}}\,, \tag{2.16}$$

where

$$\phi_{\boldsymbol{\alpha}} := a_k\cdot\hat{\phi}_{\boldsymbol{\alpha}}\,; \quad \psi_{\boldsymbol{\alpha}} := b_k\cdot\hat{\psi}_{\boldsymbol{\alpha}}\,; \quad \forall\boldsymbol{\alpha}:\ 2k-1\le|\boldsymbol{\alpha}|\le 2k\,; \quad k\in\mathbb{N}\,.$$

Fix $t\in\mathbb{N}$ arbitrary. As $\mathbf{M}_t(\boldsymbol{\phi}^{(n)})\succeq 0$ for all $n\ge t$, then by (2.16), $0\preceq\mathbf{M}_t(\boldsymbol{\phi})\preceq\mathbf{M}_t(\mu)$, and as t was arbitrary, $0\preceq\mathbf{M}_n(\boldsymbol{\phi})\preceq\mathbf{M}_n(\mu)$ for all n, and similarly $0\preceq\mathbf{M}_n(\boldsymbol{\psi})\preceq\mathbf{M}_n(\nu)$ for all n.

Next, as $\mathbf{M}_n(\boldsymbol{\phi})\preceq\mathbf{M}_n(\mu)$, and $\boldsymbol{\mu}$ satisfies Carleman's condition, then so does $\boldsymbol{\phi}$, and as $\mathbf{M}_n(\boldsymbol{\phi})\succeq 0$ for all n, it follows that $\boldsymbol{\phi}=(\phi_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^d}$ has a representing measure ϕ on $\mathbb{R}^d$. Similarly, for same reasons, $\boldsymbol{\psi}$ has a representing measure ψ on $\mathbb{R}^d$.

In addition , by (2.16),

$$\|\mu-\nu\|_{TV} \ge \lim_{k\to\infty}\rho_{n_k} = \lim_{k\to\infty}\phi^{(n_k)}(1)+\psi^{(n_k)}(1) = \phi(1)+\psi(1)\,,$$

and

$$\forall\boldsymbol{\alpha}\in\mathbb{N}^d: \quad \mu_{\boldsymbol{\alpha}}-\nu_{\boldsymbol{\alpha}} = \lim_{k\to\infty}\phi_{\boldsymbol{\alpha}}^{(n_k)}-\psi_{\boldsymbol{\alpha}}^{(n_k)} = \phi_{\boldsymbol{\alpha}}-\psi_{\boldsymbol{\alpha}}\,.$$

Hence (ϕ,ψ) is an optimal solution of (2.7) (hence of (2.5) as well), and by Corollary 2.6, $(\phi,\psi)=(\phi_+^*,\phi_-^*)$, the Hahn-Jordan decomposition of $\mu-\nu$.

Finally, as the $(n_k)_{k\in\mathbb{N}}$ was an arbitrary converging subsequence and the limit is independent of the subsequence, the whole sequence converges. □

If μ and ν are two probability measures, mutually singular, then $\|\mu-\nu\|_{TV}$=2. A perfect case to check whether (2.9) is efficient, is to test (2.9) with the toy univariate example where $\mu=\delta_0$ and $\nu=\delta_\varepsilon$ for small value of $\varepsilon>0$. Indeed, one might expect that the convergence $\rho_n\uparrow\|\mu-\nu\|_{TV}$ as n grows, could depend on ε (the smaller ε, the slower the convergence), or suffer from some numerical difficulties for small $\varepsilon>0$.

Example 1. *Let $d=1$ and $\mu=\delta_0$, $\nu=\delta_\varepsilon$, $\varepsilon\neq 0$. For this toy example we know that $\|\mu-\nu\|_{TV}=2$ and $(\phi_+^*,\phi_-^*)=(\mu,\nu)$. The semidefinite relaxation* (2.9) *with $n=1$ reads:*

$$\begin{array}{ll}\min\limits_{\boldsymbol{\phi},\boldsymbol{\psi}} & \{\,\phi_0+\psi_0:\phi_0=\psi_0\,;\ \phi_1-\psi_1=-\varepsilon\,;\ \phi_2-\psi_2=-\varepsilon^2\\ & 0\preceq\begin{bmatrix}\phi_0&\phi_1\\ \phi_1&\phi_2\end{bmatrix}\preceq\begin{bmatrix}1&0\\0&0\end{bmatrix};\ 0\preceq\begin{bmatrix}\psi_0&\psi_1\\ \psi_1&-\psi_2\end{bmatrix}\preceq\begin{bmatrix}1&\varepsilon\\ \varepsilon&\varepsilon^2\end{bmatrix}\ \}\,.\end{array}$$

The constraint $0\preceq\mathbf{M}_1(\boldsymbol{\phi})\preceq\mathbf{M}_n(\delta_0)$ yields $1\ge\phi_0$, and $\phi_2=0$, which in turn yields $\phi_1=0$. Hence $\psi_1=a$ and $\psi_2=a^2$. That is, the first semidefinite relaxation (2.9) *provides the optimal solution (ϕ_+^*,ϕ_-^*), no matter how close is ε to 0. We can see that crucial for the relaxations* (2.9) *are the domination constraints $\mathbf{M}_n(\phi)\preceq\mathbf{M}_n(\mu)$ and $\mathbf{M}_n(\psi)\preceq\mathbf{M}_n(\nu)$ (whereas they are not needed in the LP* (2.5)*).*

3. Conclusion

We have provided a numerical scheme to approximate as closely as desired the total variation distance between two measures μ and ν on $\mathbb{R}^d$. To the best of the author's knowledge, this is the first systematic algorithmic procedure to address this problem under fairly general assumptions on μ and ν, as both to satisfy Carleman's condition or the easier to check sufficient condition (2.3). On the other hand, this procedure is still "ideal" as for convergence it requires to have access to all moments of μ and ν exactly, or at least a sufficiently large finite number of them to obtain a good approximation, which can be questionable in real applications .

References

[1] *Handbook on Semidefinite, Conic and Polynomial Optimization*, Internat. Ser. Oper. Res. Management Sci., vol 166, M. Anjos and J. B. Lasserre (eds.) Springer, New York, 2012.

[2] A. Chambolle. An algorithm for total variation minimization and applications, J. Math. Imag. Vision 20, pp. 89–97, 2004.

[3] T. F. Chan, G. H. Glob, P. Mulet. A nonlinear primal-dual method for total variation-based image restoration, SIAM J. Sci. Comput. 20 (6), pp. 1964–1977, 1999.

[4] D. Henrion, M. Korda, J.B. Lasserre.*The Moment-SOS Hierarchy: Lectures in Probability, Statistics, Computational Geometry, Control and Nonlinear PDEs*, World Scientific, Singapore, 2020.

[5] D. Henrion, J. B. Lasserre. Graph recovery from incomplete moment information, Constr. Approx. 56, pp. 165–187, 2022.

[6] D. Henrion, J. B. Lasserre, J. Lofberg. Gloptipoly 3: moments, optimization and semidefinite programming, Optim. Methods and Softwares 24, pp. 761–779, 2009.

[7] J. B. Lasserre. The Moment-SOS Hierarchy, in *Proceedings of the International Congress of Mathematicians (ICM 2018)*, vol 4, B. Sirakov, P. Ney de Souza and M. Viana (eds.), World Scientific, 2019, pp. 3773–3794.

[8] J. B. Lasserre. *Moments, Positive Polynomials and Their Applications*, Imperial College Press, London, UK, 2009.

[9] M. Markatou, Yang Chen. Non-quadratic distances in model assessment, Entropy 20 (6), 464, 2018.

[10] G. Peyré, M. Cuturi. Computational Optimal Transport: With applications to Data Sscience, Found. Trends in Machine Learning 11 (5-6), pp. 355–607, 2019.

[11] L. I. Rudin, S. Osher, E. Fatemi. Nonlinear total variation based noise removal algorithms, Physica D 60, pp. 259–268, 1992.

[12] T. Weisser, B. Legat, C. Coey, L. Kapelevich, J.P. Vielma. Polynomial and moment optimization in Julia and Jump, Juliacon, 2019.

LAAS-CNRS and Toulouse School of Economics (TSE), University of Toulouse, LAAS, 7 avenue du Colonel Roche, BP 54200, 31031 Toulouse Cédex 4, France

Email address: lasserre@laas.fr