ACADÉMIE
DES SCIENCES
INSTITUT DE FRANCE

Comptes Rendus

Mathématique

Jean B. Lasserre

Gaussian mixtures closest to a given measure via optimal transport

Volume 362 (2024), p. 1455-1473

Online since: 14 November 2024

https://doi.org/10.5802/crmath.657

ACADÉMIE
DES SCIENCES
INSTITUT DE FRANCE

Comptes Rendus. Mathématique
2024, Vol. 362, p. 1455-1473
https://doi.org/10.5802/crmath.657

Research article / *Article de recherche*
Statistics / *Statistiques*

Gaussian mixtures closest to a given measure via optimal transport

Jean B. Lasserre [a]

[a] LAAS-CNRS and Toulouse School of Economics (TSE), BP 54200, 7 Avenue du Colonel Roche,
31031 Toulouse cédex 4, France

E-mail: lasserre@laas.fr

Abstract. Given a determinate (multivariate) probability measure μ, we characterize Gaussian mixtures ν_ϕ which minimize the Wasserstein distance $W_2(\mu,\nu_\phi)$ to μ when the mixing probability measure ϕ on the parameters $(\mathbf{m},\Sigma)$ of the Gaussians is supported on a compact set S. (i) We first show that such mixtures are optimal solutions of a particular optimal transport (OT) problem where the marginal ν_ϕ of the OT problem is also unknown via the mixing measure variable ϕ. Next (ii) by using a well-known specific property of Gaussian measures, this optimal transport is then viewed as a Generalized Moment Problem (GMP) and if the set S of mixture parameters $(\mathbf{m},\Sigma)$ is a basic compact semi-algebraic set, we provide a "mesh-free" numerical scheme to approximate as closely as desired the optimal distance by solving a hierarchy of semidefinite relaxations of increasing size. In particular, we neither assume that the mixing measure is finitely supported nor that the variance is the same for all components. If the original measure μ is not a Gaussian mixture with parameters $(\mathbf{m},\Sigma)\in S$, then a strictly positive distance is detected at a finite step of the hierarchy. If the original measure μ is a Gaussian mixture with parameters $(\mathbf{m},\Sigma)\in S$, then all semidefinite relaxations of the hierarchy have same zero optimal value. Moreover if the mixing measure is atomic with finite support, its components can sometimes be extracted from an optimal solution at some semidefinite relaxation of the hierarchy when Curto & Fialkow's flatness condition holds for some moment matrix.

Résumé. Étant donné une mesure de probabilité (multivariée) μ nous caractérisons les mélanges de Gaussiennes ν_ϕ qui minimisent la distance de Wasserstein $W_2(\mu,\nu_\phi)$ quand la probabilité de mélange est sur un compact S. (i) On montre d'abord que de telles probabilités de mélange sont solutions optimales d'un problème de transport où la marginale (ν_ϕ) est elle-même une inconnue via la probabilité de mélange ϕ. (ii) Ensuite en utilisant une propriété bien connue des Gaussiennes, ce problème de transport est lui-même vu comme un problème de moments généralisé. Si l'ensemble S des paramètres admissibles est un semi-algébrique de base compact, alors on fournit un schéma numérique sans grille de discrétisation (la hiérarchie « moments – sommes-de-carrés »), pour approximer arbitrairement près la distance minimum optimale. Si la mesure μ n'est pas un mélange de Gaussiennes (avec paramètres dans S) alors une distance strictement positive est détectée à une certaine relaxation de la hiérarchie. Si μ est un mélange (fini) de Gaussiennes, alors une mesure de mélange atomique peut parfois être extraite de la solution optimale d'une relaxation quand la condition de « flatness » de Curto& Fiakow est satisfaite pour une matrice de moments.

Keywords. Gaussian mixtures, Wasserstein distance, semidefinite relaxations, Moment-SOS hierarchy.

Mots-clés. Mélanges de Gaussiennes, distance de Wasserstein, relaxations semidéfinies, hiérarchie moments-sommes-de-carrés.

2020 Mathematics Subject Classification. 42C05, 47B32, 33C47, 90C23, 90C46.

Funding. J.B. Lasserre 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.

Manuscript received 24 December 2023, revised 23 April 2024, accepted 13 June 2024.

ISSN (electronic): 1778-3569

1. Introduction

Comparing mixture distributions (e.g. their "distance" to each other) is becoming an important topic with many real world applications, and particularly in data science. In addition, in the latter context, for model interpretability the mixing measure of components can be as important as the mixture distribution itself. Quoting [5], *"standard distances (Hellinger, Total Variation, Wasserstein) between mixture distributions do not capture the possibility that similar distributions may arise from mixing completely different mixture components, and have therefore different mixing measures"*. The relations between mixture distributions and their mixing measures was investigated in [21]. So for instance, in the context of *topic models*, in [5] the authors define what they call the Sketched Wasserstein Distance (SWD) between two mixture distributions, both of which consist of a finite mixing of distributions in some set of probability measures on a (Polish) space. They show that the SWD distance equals the Wasserstein distance between the mixing measures.

Among mixture distributions, Gaussian mixtures form an important subfamily because they can approximate continuous probability densities quite well. In particular they are used in statistics for clustering of data and to approximate a large family of distributions of interest in applications; see e.g. [2, 3, 6, 17–19, 23, 25, 27]. Mixtures of Gaussians $\mathcal{N}(\mathbf{m},\Sigma)$ on $\mathbb{R}^d$ have the well-known and nice property that every moment $\mu_{\boldsymbol{\alpha}} = \int \mathbf{x}^{\boldsymbol{\alpha}}\,d\mu$, $\boldsymbol{\alpha}\in\mathbb{N}^d$, is an *explicit* polynomial of degree $|\boldsymbol{\alpha}|$ in the parameters $(\mathbf{m},\Sigma)$ of the mixture, and therefore determining whether a real sequence $(y_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^d}$ has a representing measure μ which is some Gaussian mixture, has been recently investigated in e.g. [2, 3] as a specific moment-problem in real analysis. In particular in [3] the authors prove positive and negative results on rational identifiability[1] of k-atomic mixing measures of mixture distributions; for instance if $d=1$ then for all k, a k-atomic mixing measure can be identified from sufficiently many moments of the mixing distribution [3, Theorem 1]. The same result for mixtures of bivariate Gaussians is a conjecture [3, Conjecture 2]; see also [16] on the key role of moment matrices and determinants in the method of moments.

On the other hand, an important problem in robust statistics is to estimate parameters of Gaussian mixtures from their samples (possibly with noisy data). In contributions [4, 11, 12] from the theoretical computer science community, (theoretical) polynomial time algorithms (e.g. sum-of-squares algorithms) have been proposed for efficient learning of mixtures with asymptotic guarantees. In the recent contribution [26], a practical algorithm for optimal estimation of mixtures of finitely many univariate Gaussians with same (known or unknown) variance is proposed via a (denoised) method of moments. It combines semidefinite programming and Gauss quadratures to estimate a mixture of k univariate Gaussians with same variance. In [8] the authors consider the estimator made of mixtures with k atoms (and same variance) which minimizes the Kolmogorov distance of its distribution function to that of the input distribution, and they provide optimal rates of estimation (the k-atomic mixing distributions are compared with the Wasserstein distance) but no algorithm is provided. Again, the notion of k-idenfiability is of central importance in [8].

In this paper we consider the following problem: Given a probability measure μ on $\mathbb{R}^d$, and a compact set S of parameters $(\mathbf{m},\Sigma)$, find a mixture ν of Gaussian measures $\mathcal{N}(\mathbf{m},\Sigma)$ with with parameters $(\mathbf{m},\Sigma)\in S$, which is the closest to μ. How close is ν to μ is measured e.g. by the 2-Wasserstein (or Kantorovich) distance $W_2(\mu,\nu)$. That is:

$$\nu(B) = \int_S \left(\frac{1}{(2\pi)^{d/2}\sqrt{\det(\Sigma)}}\int_B \exp(-(\mathbf{x}-\mathbf{m})^T\Sigma^{-1}(\mathbf{x}-\mathbf{m})/2)\,d\mathbf{x}\right) d\phi(\mathbf{m},\Sigma)\,, \quad \forall B\in\mathscr{B}(\mathbb{R}^d)\,,$$

[1] Algebraic identifiability means that there are finitely many (complex) solutions to the moment equations for generic values of the sample moments. On the other hand, rational identifiability is about generic uniqueness of real solutions, up to the label-swapping action of the symmetric group S_k

for some probability ϕ on S (the mixing measure of parameters $(\mathbf{m},\Sigma)\in S$), and

$$W_2(\mu,\nu)^2=\inf_{\lambda}\left\{\int_{\mathbb{R}^{2d}}\|\mathbf{x}-\mathbf{y}\|^2\,\mathrm{d}\lambda(\mathbf{x},\mathbf{y}):\,\lambda_{\mathbf{x}}=\mu;\,\lambda_{\mathbf{y}}=\nu\right\},$$

where λ is a probability measure on $\mathbb{R}^{2d}$, and $\lambda_{\mathbf{x}}$ (resp. $\lambda_{\mathbf{y}}$) denotes the marginal of λ w.r.t. $\mathbf{x}$ (resp. w.r.t. $\mathbf{y}$). In fact, the results and proposed methodology are also valid if one uses the 1-Wasserstein distance W_1 instead of W_2.

Statement of the problem and contribution

For sake of clarity and simplicity of exposition, we first restrict to the univariate case. Then we briefly describe extension to the multivariate case. While this extension does not pose any theoretical problem, on the other hand the associated numerical scheme is more demanding (simply for question of scalability of the approach).

Statement of the problem. Let $\mathscr{P}(\mathscr{X})$ denote the space of probability measures on a Borel set $\mathscr{X}\subset\mathbb{R}^2$. With $\mathbb{R}_+:=\{x:x\geq 0\}$, let $S\subset\mathbb{R}\times\mathbb{R}_+$ be a set of parameters (m,σ) for univariate Gaussian measures $\mathscr{N}(m,\sigma)$, and let $\boldsymbol{\mu}=(\mu_j)_{j\in\mathbb{N}}$ be the moment sequence of a given probability measure μ on the real line. The goal is to find a Gaussian mixture ν with mixing parameters in S that is the closest to μ with respect to the Wasserstein distance

$$W_2(\mu,\nu)^2=\min_{\lambda\in\mathscr{P}(\mathbb{R}^2)}\left\{\int_{\mathbb{R}^2}(x-y)^2\,\mathrm{d}\lambda(x,y):\,\lambda_x=\mu;\,\lambda_y=\nu\right\},\tag{1}$$

where λ_x (resp. λ_y) is the marginal of λ w.r.t. x (resp. w.r.t. y) on $\mathbb{R}$. Alternatively one may also use the Wasserstein distance $W_1(\mu,\nu)=\int|x-y|\,\mathrm{d}\lambda$ (see Appendix).

As ν is required to be a Gaussian mixture, it is associated with some (not necessarily unique) *mixing* probability measure ϕ on the set S of Gaussian parameters (m,σ), and therefore ν is in fact denoted by ν_ϕ, and reads

$$\nu_\phi(B):=\int_S\left(\frac{1}{\sqrt{2\pi}\sigma}\int_B\exp\left(\frac{-(x-m)^2}{2\sigma^2}\right)\mathrm{d}x\right)\mathrm{d}\phi(m,\sigma),\quad\forall B\in\mathscr{B}(\mathbb{R}).\tag{2}$$

Equivalently, ν_ϕ has the density

$$x\longmapsto\int_S\frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-(x-m)^2}{2\sigma^2}\right)\mathrm{d}\phi(m,\sigma),$$

w.r.t. Lebesgue measure on $\mathbb{R}$. Therefore one wishes to solve the optimization problem

$$\tau=\inf_{\phi\in\mathscr{P}(S)}W_2(\mu,\nu_\phi)^2=\inf_{\phi\in\mathscr{P}(S),\lambda\in\mathscr{P}(\mathbb{R}^2)}\left\{\int(x-y)^2\,\mathrm{d}\lambda(x,y):\,\lambda_x=\mu;\,\lambda_y=\nu_\phi\right\}.\tag{3}$$

Observe that (3) is an optimal transport problem of a particular type. Indeed the second marginal $\lambda_y=\nu_\phi$ of the unknown λ is also to be optimized via the (mixing measure) variable ϕ on S.

Contribution. We assume that the set of parameters $S\subset\mathbb{R}\times\mathbb{R}_+$ is compact. In contrast to previous works we do *not* assume that the mixing measure is k-atomic (and not even with same variance for all components). Also our algorithm is potentially and directly applicable to mixtures of multivariate Gaussians, although of course its efficiency strongly depends on the dimension. At last, the input probability measure μ is not necessarily a Gaussian mixture and our primary goal is to evaluate how far is μ from a mixture of Gaussians with parameters (m,σ) in a given set S. If μ is indeed such a Gaussian mixture then the algorithm helps to detect an associated mixing measure.

(I) We first show that if μ satisfies

$$\int \exp(c\,|x|)\,\mathrm{d}\mu(x) < \infty\,, \tag{4}$$

for some scalar $c > 0$, then (3) has an optimal solution $(\lambda^*,\phi^*) \in \mathscr{P}(\mathbb{R}^2)\times\mathscr{P}(S)$ (i.e., $\tau = W_2(\mu,\nu_{\phi^*})^2$). Moreover, introducing the moment sequences $\boldsymbol{\lambda}^* = (\lambda^*_{(i,j)})_{(i,j)}$ and $\boldsymbol{\phi}^* = (\phi^*_{(i,j)})_{(i,j)}$, with

$$\lambda^*_{(i,j)} = \int x^i y^j\,\mathrm{d}\lambda\,,\quad \phi^*_{(i,j)} = \int m^i\sigma^j\,\mathrm{d}\phi^*\,,\quad \forall (i,j)\in\mathbb{N}^2\,,$$

the couple (λ^*,ϕ^*) is also an optimal solution of:

$$\inf_{\lambda\in\mathscr{P}(\mathbb{R}^2),\phi\in\mathscr{P}(S)} \left\{ \int (x-y)^2\,\mathrm{d}\lambda : \lambda_{(j,0)} = \mu_j\,;\quad \lambda_{(0,j)} = \int p_j(m,\sigma)\,\mathrm{d}\phi,\quad \forall j\in\mathbb{N}\right\}, \tag{5}$$

which is an *exact moment-relaxation* of (3). To show that (5) is equivalent to (3), one exploits that (i) S is compact, (ii) the well-known fact that every moment μ_j of a Gaussian measure $\mu = \mathscr{N}(m,\sigma)$ is an explicit polynomial $p_j \in \mathbb{R}[m,\sigma]$ of degree j, and (iii) that μ is moment determinate (because of (4)). To the best of our knowledge, this is the first characterization of best Wasserstein-approximations by Gaussian mixtures (with parameters in a given set S) as *optimal solutions of an optimal transport problem.*

We also obtain that strong duality holds between (5) and its dual which reads

$$\sup_{q\in\mathbb{R}[x],g\in\mathbb{R}[y]} \left\{ \int q\,\mathrm{d}\mu : q(x)+g(y) \le (x-y)^2,\ \forall x,y; \right.$$
$$\left. \frac{1}{\sqrt{2\pi}\sigma}\int g(x)\,\exp\left(\frac{-(x-m)^2}{2\sigma^2}\right)\mathrm{dx} \ge 0,\quad \forall (m,\sigma)\in S\right\}, \tag{6}$$

and is very close in spirit to the classical dual of the Monge-Kantorovich optimal transport (with cost $\|\mathbf{x}-\mathbf{y}\|^2$).

(II) Next, the exact moment formulation (5) of (3) is a particular instance of the "Generalized Moment Problem" (GMP) (see e.g. [13]) whose description is trough algebraic data only (because every moment of a Gaussian is a *polynomial* in the parameters (m,σ)). Therefore one can apply the *Moment-SOS hierarchy* [9, 13] to solve (5). That is, the optimal value τ of (5) (hence of (3) as well) can be approximated as closely as desired by solving a sequence (a hierarchy) of semidefinite relaxations of increasing size (as more and more moments are taken into account).

The degree-n semidefinite relaxation of (3) (and of (5)) is just (5) where $\phi\in\mathscr{P}(S)$ and $\lambda\in\mathscr{P}(\mathbb{R}^2)$ are respectively replaced with degree-$2n$ pseudo-moment sequences $\boldsymbol{\phi} = \phi_{(i,j)})_{(i,j)\in\mathbb{N}^2_{2n}}$ and $\boldsymbol{\lambda} = (\lambda_{(i,j)})_{(i,j)\in\mathbb{N}^2_{2n}}$, that satisfy necessary semidefinite constraints to be moments of a measure on S and $\mathbb{R}^2$ respectively, coming from Putinar's Positivstellensatz [13, 24].

If the input measure is not a mixture of Gaussians with parameters $(m,\sigma)\in S$, then the optimal value becomes strictly positive at some step of the hierarchy, which provides a certificate that μ cannot be a mixture of Gaussians with parameters $(m,\sigma)\in S$ (i.e., of the form (2)).

(III) On the other hand, if the input measure μ *is* a mixture of finitely many Gaussian measures with parameters $(m,\sigma)\in S$, then $\tau = 0$, $\lambda^* = \mu\otimes\mu$, and ϕ^* is an atomic mixing measure (not necessarily unique) with finite support. If a certain rank condition (Curto & Fialkow's flat extension in [13, Theorem 3.11]) is satisfied at an optimal solution $(\widehat{\boldsymbol{\lambda}},\widehat{\boldsymbol{\phi}})$ of some degree-n relaxation in the hierarchy (with optimal value zero), then the support and weights of some atomic measure $\widehat{\phi}$ on S can be recovered from $\widehat{\boldsymbol{\phi}}$. To check whether

$\widehat{\phi}$ is optimal for (5) (and $\widehat{\phi} = \phi^*$ and ϕ^* is unique) can be done by checking whether all moments of $\nu_{\hat{\phi}}$ of degree higher than $n+1$ match those of μ, i.e., whether

$$\mu_j = \int p_j(m,\sigma)\,\mathrm{d}\widehat{\phi}(m,\sigma)\,, \quad \forall j > n+1\,. \tag{7}$$

Checking (7) for each fixed $j > n+1$ is easy and can be done exactly.

We recall that identifiability of the mixing measure from moments of the mixture distribution is a delicate issue [3] as in general, several mixing measures can be solutions. However in our setting we have the additional condition that the mixing mesure is supported on S.

Again we emphasize our minimal assumptions: the input measure μ satisfies (4) and the parameter set S of admissible mixtures of Gaussians is a compact basic semi-algebraic set. In particular and in contrast to [26], the variance σ is not fixed and the mixing measures are not assumed to be atomic with finite support.

The paper closest in spirit to ours is the practical algorithm [26] for mixtures μ of k *univariate* Gaussian measures with *same* variance σ (both cases where σ is known and unknown are considered in [26]). The author first estimates a vector of $2k-1$ moments of μ via Hermite polynomials, then denoises this vector by projection onto the moment space (via semidefinite programming), and then obtains a resulting k-atomic distribution via Gauss quadrature. Nice results in [26, Theorem 1; (8)] provide optimal rates (with respect to Wasserstein distance W_1) provided that k and σ are known, and [26, Theorem 1; (9)-(10)] if k is known whereas σ is unknown. In [26] the semidefinite program is used to “denoise” the input vector of moments by projection onto the moment space. The Wasserstein distance is only used to quantify *a posteriori* the error and justify convergence. In our approach, the semidefinite relaxation (i) models directly the Wasserstein distance W_2 (using W_1 is also possible) between the input measure and any Gaussian mixture ν_ϕ, and (ii) is parametrized by the number of moments considered. Finally, notice that the approach in [26] is possible thanks to very specific features that are proper to the univariate case only. Namely:

- (convex) semidefinite programming constraints (exploited in [26]) provide *necessary and sufficient* conditions for a finite real sequence to have a representing measure and so the output of the semidefinite program in [26] is a true moment sequence; but similar conditions are only necessary in the multivariate setting.
- similarly, Gauss quadratures also exploited in [26] do not always exist in the multivariate setting (then called Gauss cubatures); see e.g. [7, 14, 20].

For ease and clarity of exposition, we concentrate in the univariate case but all results of Section 3 are also extended to the multivariate case which is briefly addressed in Section 4.

2. Notation, definitions and preliminary results

2.1. *Notation and definitions*

Let $\mathbb{R}[x,y]$ denote the ring of real polynomials in the two variables (x,y) and $\mathbb{R}[x,y]_n \subset \mathbb{R}[x,y]$ be its subset of polynomials of total degree at most n. Let $\mathbb{N}^2_n := \{(i,j) \in \mathbb{N}^2 : i+j| \leq n\}$ with cardinal $s(n) = \binom{n+2}{2}$. Let $\mathbf{v}_n(x,y) = (x^i y^j)_{(i,j)\in\mathbb{N}^2_n}$ be the vector of monomials up to degree n, and let $\Sigma[x,y]_n \subset \mathbb{R}[x,y]_{2n}$ be the convex cone of polynomials of total degree at most $2n$ which are sum-of-squares (in short SOS). A polynomial $p \in \mathbb{R}[x,y]_n$ can be identified with its vector of coefficients $\mathbf{p} = (p_{(i,j)}) \in \mathbb{R}^{s(n)}$ in the monomial basis, and reads

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

With $\mathscr{X}\subset\mathbb{R}^2$, denote by $\mathscr{M}(\mathscr{X})_+$ (resp. $\mathscr{C}(\mathscr{X})$), the space of positive measures (resp. continuous functions) on $\mathscr{X}$, and by $\mathscr{P}(\mathscr{X})$, the space of probability measures on $\mathscr{X}$.

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.)). The support of a Borel measure μ on $\mathbb{R}^2$ is the smallest closed set A such that $\mu(\mathbb{R}^2\setminus A)=0$, and such a set A is unique. A Borel measure with all moments finite is said to be (moment) *determinate* if there is no other measure with same moments.

Riesz functional, moment and localizing matrix. With a real sequence $\boldsymbol{\phi}=(\phi_{(i,j)})_{(i,j)\in\mathbb{N}^2}$ (in bold) is associated the *Riesz* linear functional $\phi\in\mathbb{R}[x,y]^*$ (not in bold) defined by

$$p\left(=\sum_{(i,j)}p_{i,j}x^iy^j\right)\longmapsto\phi(p)=\langle\boldsymbol{\phi},\mathbf{p}\rangle=\sum_{\boldsymbol{\alpha}}p_{i,j}\,\phi_{(i,j)},\quad\forall p\in\mathbb{R}[x,y],$$

and the moment matrix $\mathbf{M}_n(\boldsymbol{\phi})$ with rows and columns indexed by $\mathbb{N}^2_n$ (hence of size $s(n)$), and with entries

$$\mathbf{M}_n(\boldsymbol{\phi})((i,j),(i',j')):=\phi(x^{i+i'}y^{j+j'})=\phi_{(i+i',j+j')},\quad(i,j),(i',j')\in\mathbb{N}^2_n.$$

Similarly, given $g\in\mathbb{R}[x,y]$ ($(x,y)\mapsto\sum_{(i,j)}g_{i,j}x^iy^j$), define the new sequence

$$g\cdot\boldsymbol{\phi}:=\left(\sum_{(k,\ell)}g_{k,\ell}\,\phi_{(i,j)+(k,\ell)}\right)_{(i,j)\in\mathbb{N}^2},$$

and the localizing matrix associated with $\boldsymbol{\phi}$ and g,

$$\mathbf{M}_n(g\cdot\boldsymbol{\phi})((i,j),(i',j')):=\sum_{(k,\ell)}g_{k,\ell}\,\phi_{(i+i'+k,j+j'+\ell)},\quad(i,j),(i',j')\in\mathbb{N}^2_n.$$

Equivalently, $\mathbf{M}_n(g\cdot\boldsymbol{\phi})$ is the moment matrix associated with the new sequence $g\cdot\boldsymbol{\phi}$. The Riesz linear functional $g\cdot\phi$ associated with the sequence $g\cdot\boldsymbol{\phi}$ satisfies

$$g\cdot\phi(p)=\phi(g\,p),\quad\forall p\in\mathbb{R}[x,y].$$

A real sequence $\boldsymbol{\phi}=(\phi_{(i,j)})_{(i,j)\in\mathbb{N}^2}$ has a representing mesure if its associated linear functional ϕ is a Borel measure on $\mathbb{R}^2$. In this case $\mathbf{M}_n(\boldsymbol{\phi})\succeq 0$ for all n; the converse is not true in general. In addition, if ϕ is supported on the set $\{(x,y)\in\mathbb{R}^2:g(x,y)\geq 0\}$ then $\mathbf{M}_n(g\cdot\boldsymbol{\phi})\succeq 0$ for all n.

Multivariate Carleman condition. The following condition due to Carleman in the univariate case and later extended by Nussbaum to the multivariate case, is a very useful sufficient condition to ensure that a moment sequence has a representing measure; see e.g. [13, Theorem 3.13]. We here specialize to the 2-dimensional case.

Theorem 1 (Bivariate Carleman condition). *Let $\boldsymbol{\phi}=(\phi_{(i,j)})_{(i,j)\in\mathbb{N}^2}$ be a real sequence such that $\mathbf{M}_n(\boldsymbol{\phi})\succeq 0$ for all n, and such that*

$$\sum_{j=1}^{\infty}(\phi_{(2j,0)})^{-1/2j}=+\infty;\quad\sum_{j=1}^{\infty}(\phi_{(0,2j)})^{-1/2j}=+\infty.\tag{8}$$

Then $\boldsymbol{\phi}$ has a representing measure ϕ on $\mathbb{R}^2$ and ϕ is moment determinate.

For instance, if ϕ is a finite Borel measure on $\mathbb{R}^2$ and $\sup[\int\exp(c|x|)\,d\phi,\int\exp(c'|y|)\,d\phi]<\infty$ for some scalars $c,c'>0$, then the moment sequence $\boldsymbol{\phi}$ satisfies (8), and ϕ is moment determinate.

2.2. *An intermediate result*

The following result is well-known and is reproduced for sake of clarity.

Proposition 2. *If $\sigma > 0$ then for every $j \in \mathbb{N}$, the moment*

$$(m,\sigma) \longmapsto \frac{1}{\sqrt{2\pi}\sigma}\int x^j \exp\frac{-(x-m)^2}{2\sigma^2}\,\mathrm{dx}, \tag{9}$$

is a polynomial $p_j \in \mathbb{R}[m,\sigma]$ of total degree at most j, and:

$$p_{2j}(m,\sigma) = \sum_{k=0}^{j}(2k-1)!!\sigma^{2k}\,m^{2(j-k)}\binom{2j}{2k}, \quad \forall j \in \mathbb{N}. \tag{10}$$

Moreover, if $\sigma = 0$ then

$$p_{2j}(m,0) = m^{2j} = \int x^{2j}\,\delta_m(\mathrm{dx}), \quad \forall j \in \mathbb{N}. \tag{11}$$

Proof. Recall that

$$\frac{1}{\sqrt{2\pi}\sigma}\int (x-m)^j \exp\frac{-(x-m)^2}{2\sigma^2}\,\mathrm{dx} = \begin{cases} 0 & \text{if } j \text{ is odd,} \\ \sigma^j(j-1)!! & \text{if } j \text{ is even,} \end{cases} \quad \forall j \in \mathbb{N}, \tag{12}$$

with for $j \geq 2$, $j!! = j(j-2)(j-4)\cdots$, $1!! = 1$, and the convention $-1!! = 1$. For instance, $p_0 = \mathbf{1}$, $p_1(m,\sigma) = m$, $p_2(m,\sigma) = m^2 + \sigma^2$, etc. Next, doing the change of variable $u = (x-m)$ in the integrand of (10), expanding $(u+m)^j$ in the basis of monomials, and summing up, yields (10). □

Remarks 3.

(i) A Gaussian mixture is associated with a (non necessarily unique) mixing probability $\phi \in \mathscr{P}(S)$ and in view of (11), ϕ may tolerate that $\phi(\{\mathbb{R}\times\{0\}) > 0$, i.e., ϕ can mix Gaussian densities with discrete measures. In other words and with a slight abuse of notation, the Dirac measure δ_m at point m can be viewed a the degenerate "Gaussian measure" $\mathscr{N}(m,0)$, with vector of moments $(m^j)_{j\in\mathbb{N}} = (p_j(m,0))_{j\in\mathbb{N}}$. For instance if $\mu = \sum_{k=1}^{s}\gamma_k\delta_{x_k}$ for some set $\{x_1,\ldots,x_k\} \subset \mathbb{R}$ and scalars $\gamma_k \geq 0$, i.e., a mixture of s Dirac measures with weights (γ_k), then

$$\mu_j = \int x^j\,\mathrm{d}\mu = \sum_{k=1}^{s}\gamma_k\,x_k^j = \sum_{k=1}^{s}\gamma_k\,p_j(x_k,0) =: \int x^j\left(\sum_{k=1}^{s}\gamma_k\,\mathrm{d}\mathscr{N}(x_k,0)\right), \quad \forall j \in \mathbb{N}.$$

(ii) So as a consequence, if $S = [-M,M]\times[0,\overline{\sigma}]$ then every measure μ on $[-M,M]$ can be considered a Gaussian mixture where μ itself is the mixing measure. Indeed its moments $(\mu_j)_{j\in\mathbb{N}}$ satisfy

$$\mu_j = \int m^j\,\mathrm{d}\mu(m) = \int p_j(m,0)\,\mathrm{d}\mu(m) = \int\left(\int x^j\mathrm{d}\mathscr{N}(m,0)\right)\mathrm{d}\mu(m), \quad j \in \mathbb{N}.$$

In particular, every discrete measure on $[-M,M]$ is also a Gaussian mixture with parameters $(m,0) \in S$. This is not what one usually has in mind when thinking of Gaussian mixtures, as one would expect a measure μ with a density w.r.t. Lebesgue measure on $\mathbb{R}$. So this is why one should assume that the compact set S satisfies $\sigma \geq \delta > 0$ for all $(m,\sigma) \in S$, for some positive scalar δ; for instance, $S := [-M,M]\times[\underline{\sigma},\overline{\sigma}]$ with $\underline{\sigma} > 0$.

Corollary 4. *Let ϕ be a probability measure on S. Then with $p_{2j} \in \mathbb{R}[m,\sigma]$, $j \in \mathbb{N}$, as in* (10)

$$\sum_{j=1}^{\infty}\phi(p_{2j})^{-1/2j} = +\infty. \tag{13}$$

Proof. Observe that as S is compact, there exists $M > 0$ such that $|m|, \sigma < M$ for all $(m,\sigma) \in S$, and so in particular,

$$\begin{aligned} p_{2j}(m,\sigma) < M^{2j} \sum_{k=1}^{j} \frac{(2j)!}{(2(j-k))!}\frac{(2k)!!}{(2k)!} &< \quad M^{2j} \sum_{k=1}^{j} \frac{(2j)(2j-1)\cdots(2j-(2k-1))}{(2k-1)!!} \\ &< \quad M^{2j} \sum_{k=1}^{j} (2j)^{2k-1} < M^{2j} \sum_{k=1}^{j} (2j)^{2j-1} \\ &< \quad (2Mj)^{2j}, \end{aligned} \tag{14}$$

and therefore if ϕ is a probability measure on S, then $\phi(p_{2j}) < (2Mj)^{2j}$ for all $j \in \mathbb{N}$, which in turn implies the desired result

$$\sum_{j=1}^{\infty} \phi(p_{2j})^{-1/2j} > \frac{1}{2M} \sum_{j=1}^{\infty} j^{-1} = +\infty. \tag{15}$$

□

3. Main result

3.1. *The optimal transport problem* (3) *and its exact moment relaxation* (5)

Consider the optimal transport problem (3).

Theorem 5. *Let $S \subset \mathbb{R} \times \mathbb{R}_+$ be compact, and assume that $\mu \in \mathscr{P}(\mathbb{R})$ satisfies* (4).

(i) *The optimal transport problem* (3) *has an optimal solution $(\phi^*, \lambda^*) \in \mathscr{P}(S) \times \mathscr{P}(\mathbb{R}^2)$ which is also an optimal solution of* (5). *Moreover, both measures $\lambda^* \in \mathscr{P}(\mathbb{R}^2)$ and $\nu_{\phi^*} \in \mathscr{P}(\mathbb{R})$ are moment determinate.*

(ii) *Moreover, $\tau = 0$ if and only if $\lambda^* = \mu \otimes \mu$ and $\mu = \nu_{\phi^*}$, i.e., μ is a Gaussian mixture with ϕ^* a mixing measure of parameters $(m,\sigma) \in S$.*

For clarity of exposition a proof is postponed to Section 6.

Remarks 6.

(i) Notice that the mixing probability measure $\phi^* \in \mathscr{P}(S)$ is not necessary unique. That is, two different mixing measures ϕ_1 and ϕ_2 may produce the same mixture distribution $\nu_{\phi_1} = \nu_{\phi_2}$. This uniqueness issue is related to rational identifiability issue already mentioned and explored in e.g. [2, 3]. However in our restricted setting, uniqueness is perhaps easier to get as the support of the mixing measure is *not* the whole space $\mathbb{R}^2$ but a compact set $S \subset \mathbb{R}^2$.

(ii) In Theorem 5, ν_{ϕ^*} is a mixture of Gaussian measures with parameters $(m,\sigma) \in S$. If $\sigma = 0$ is tolerated in $(m,\sigma) \in S$, the mixture ϕ^* can be made of "pure" Gaussian measures $\mathcal{N}(m,\sigma)$ with $\sigma > 0$ and atomic measures δ_m“ = ”$\mathcal{N}(m,0)$. If one wishes to obtain the closest mixture ν_{ϕ^*} of "pure" Gaussian measures $\mathcal{N}(m,\sigma)$ with $\sigma > 0$, (i.e., with no atomic part), then in Theorem 5 one should replace $S \subset \mathbb{R} \times \mathbb{R}_+$ with $S \subset \mathbb{R} \times \mathbb{R}_{++}$ (with $\mathbb{R}_{++} := \{x : x > 0\}$). As S is assumed to be compact this implies that for some $\delta > 0$, $\sigma \geq \delta$ for all $(m,\sigma) \in S$.

The interesting case is precisely when $\sigma = 0$ is *not* tolerated. Indeed if $\sigma = 0$ is tolerated then any probability measure μ supported on the set $\{m : (m,0) \in S\}$ (in particular atomic measures) is the "Gaussian mixture" $\mathcal{N}(m,0)\,\mathrm{d}\mu(m)$ with mixing measure μ itself, which is not really what one wants to detect. see Remark 3(i).

A dual of (3)**.** For any $g\in\mathbb{R}[y]$ write $y\mapsto g(y):=\sum_k g_k y^k$ where (g_k) is the vector of coefficients of g in the monomial basis $(y^k)_{k\in\mathbb{N}}$. Consider the optimization problem:

$$\tau^* = \sup_{q\in\mathbb{R}[x],g\in\mathbb{R}[y]} \left\{ \int q\,\mathrm{d}\mu : \; q(x)+g(y)\leq (x-y)^2 \quad \forall x,y\in\mathbb{R}; \right.$$
$$\left. \sum_k g_k\, p_k(m,\sigma)\geq 0, \quad \forall (m,\sigma)\in S \right\}. \tag{16}$$

Observe that:

$$\sum_k g_k\, p_k(m,\sigma)\geq 0,\quad \forall (m,\sigma)\in S \Longleftrightarrow \frac{1}{\sqrt{2\pi}\sigma}\int g(x)\exp\left(\frac{-(x-m)^2}{2\sigma^2}\right)\mathrm{dx}\geq 0,\quad \forall (m,\sigma)\in S.$$

Proposition 7. *The optimization problem* (16) *is a dual of* (3)*, i.e., weak duality $\tau\geq\tau^*$ holds.*

Proof. Let (λ,ϕ) (resp. (q,g)) be a feasible solution of (3) (resp. (16)). Then as $\lambda_x=\mu$ and $\lambda_y=\nu_\phi$,

$$\int (x-y)^2\,\mathrm{d}\lambda(x,y)\geq \int (q+g)\,\mathrm{d}\lambda = \int q\,\mathrm{d}\mu + \int g\,\mathrm{d}\nu_\phi$$
$$= \int q\,\mathrm{d}\mu + \int_S \underbrace{\left(\sum_k g_k\, p_k\right)}_{\geq 0 \text{ on } S}\mathrm{d}\phi \geq \int q\,\mathrm{d}\mu,$$

and as (λ,ϕ) and (q,g) are arbitrary feasible solutions, it follows that $\tau\geq\tau^*$. □

3.2. *A hierarchy of semidefinite relaxations*

We here consider the case where the set $S\subset\mathbb{R}^2$ of parameters (m,σ) is the compact basic semi-algebraic set

$$S=\{(m,\sigma):\; u_j(m,\sigma)\geq 0,\; j=1,\ldots,s\}, \tag{17}$$

for some polynomials $u_j\subset\mathbb{R}[m,\sigma]$, $j=1,\ldots,s$, and we let $u_0:=\mathbf{1}$ (the constant polynomial equal to 1 for all (m,σ). Moreover as S is compact, we also assume that we know a scalar R such that $S\subset\{(m,\sigma): m^2+\sigma^2<R^2\}$ and without changing S we include the redundant quadratic constraint $R^2-m^2-\sigma^2\geq 0$ in its definition (17), with for instance $u_1(m,\sigma)=R^2-m^2-\sigma^2$.

Next, let $d_j:=\lceil\deg(u_j)/2\rceil$, $n_0:=\max_j d_j$ and fix $n\geq n_0$. With $p_j\in\mathbb{R}[m,\sigma]$ as in (9), define:

$$\tau_n=\min_{\boldsymbol{\phi},\boldsymbol{\lambda}}\Big\{\lambda((x-y)^2):\quad \lambda_{(j,0)}=\mu_j;\quad \lambda_{(0,j)}-\phi(p_j(m,\sigma))=0,\quad \forall j\leq 2n;$$
$$\mathbf{M}_n(\boldsymbol{\lambda})\succeq 0, \mathbf{M}_n(\boldsymbol{\phi})\succeq 0, \mathbf{M}_{n-d_j}(u_j\cdot\boldsymbol{\phi})\succeq 0,\quad j=0,\ldots,s\Big\}, \tag{18}$$

where $\boldsymbol{\lambda}=(\lambda_{(i,j)})_{(i,j)\in\mathbb{N}^2_{2n}}$ and $\boldsymbol{\phi}=(\phi_{(i,j)})_{(i,j)\in\mathbb{N}^2_{2n}}$. Problem (18) is a semidefinite program[2]. Its dual reads:

$$\tau^*_n=\sup_{q,g,\sigma,\theta_j}\left\{\int q\,\mathrm{d}\mu:\quad q(x)+g(y)+\sigma(x,y)=(x-y)^2,\quad \forall x,y\in\mathbb{R};\right.$$
$$\sum_{k=0}^{2n} g_k\, p_k(m,\sigma)=\sum_{j=0}^{s}\theta_j(m,\sigma)\,u_j(m,\sigma);$$
$$\left. q\in\mathbb{R}[x]_{2n}, g\in\mathbb{R}[y]_{2n};\, \sigma\in\Sigma[x,y]_n;\, \theta_j\in\Sigma[m,\sigma]_{n-d_j},\; j=0,\ldots,s\right\}, \tag{19}$$

with $\tau^*_n\leq\tau_n$ for all $n\geq n_0$.

[2] A semidefinite program is a convex conic program on the cone of positive semidefinite matrices. Up to arbitrary (but fixed) precision, it can be solved efficiently; see e.g. [1, 22]

Lemma 8. *For each fixed $n \geq n_0$,* (18) *is a semidefinite program and a convex relaxation of the infinite-dimensional problem* (3) *and so $\tau_n \leq \tau$ for all $n \geq n_0$. Moreover, if S has nonempty interior and* $\mathrm{supp}(\mu)$ *contains an open set, then $\tau_n = \tau_n^*$ and* (19) *has an optimal solution $(q^*, g^*, \theta_0^*, \ldots, \theta_s^*)$.*

Proof. Let $(\lambda, \phi) \in \mathscr{P}(\mathbb{R}^2) \times \mathscr{P}(S)$ be a feasible solution of (3), and let $\boldsymbol{\lambda} = (\lambda_{(i,j)})_{(i,j)\in\mathbb{N}^2_{2n}}$ and $\boldsymbol{\phi} = (\phi_{(i,j)})_{(i,j)\in\mathbb{N}^2_{2n}}$ be the vectors of degree-$2n$ moments of λ and ϕ respectively. Then the couple $(\boldsymbol{\lambda}, \boldsymbol{\phi})$ is a feasible solution of (18), and so $\tau_n \leq \tau$ for all $n \geq n_0$. Next, let ϕ be the probability measure uniformly distributed on S, and let $\lambda := \mu \otimes \nu_\phi$. Then as S has nonempty interior, $\mathbf{M}_n(u_j \cdot \boldsymbol{\phi}) \succ 0$ for all $j = 0, \ldots, s$, and $\mathbf{M}_n(\boldsymbol{\lambda}) \succ 0$. Indeed, suppose that for some $h \in \mathbb{R}[x,y]_n$ with coefficient vector $\mathbf{h}$,

$$\begin{aligned} 0 = \langle \mathbf{h}, \mathbf{M}_n(\boldsymbol{\lambda})\,\mathbf{h}\rangle &= \int h(x,y)^2\,\mathrm{d}\lambda(x,y) \\ &= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} h(x,y)^2\,\mathrm{d}\nu_\phi(y) \right) \mathrm{d}\mu(x) \\ &= \int_{\mathbb{R}} \left(\int_S \frac{1}{\sqrt{2\pi}\sigma} \int_{\mathbb{R}} h(x,y)^2 \exp(-(y-m)^2/2\sigma^2)\,\mathrm{d}y\,\mathrm{d}\phi(m,\sigma) \right) \mathrm{d}\mu(x). \end{aligned}$$

We next prove that then $h = 0$ and so $\mathbf{M}_n(\boldsymbol{\lambda}) \succ 0$. Observe that with $h \in \mathbb{R}[x,y]_n$, one may write

$$h(x,y)^2 = \sum_{k=0}^{2n} \theta^h_{n-k}(x)\, y^k, \quad \text{with } \theta^h_{n-k} \in \mathbb{R}[x]_{2n-k} \text{ for all } k = 0, \ldots, 2n,$$

and therefore

$$\frac{1}{\sqrt{2\pi}\sigma} \int_{\mathbb{R}} h(x,y)^2 \exp(-(y-m)^2/2\sigma^2)\,\mathrm{d}y =: \sum_{k=0}^{2n} \theta^h_{n-k}(x)\, p_k(m,\sigma) =: q_h(x,m,\sigma),$$

is a polynomial in $\mathbb{R}[x,m,\sigma]_{2n}$. Moreover, for all $x \in \mathbb{R}$,

$$q_h(x,m,\sigma) \geq \left(\frac{1}{\sqrt{2\pi}\sigma} \int_{\mathbb{R}} |h(x,y)| \exp(-(y-m)^2/2\sigma^2)\,\mathrm{d}y \right)^2 \geq 0, \quad \forall (m,\sigma) \in S.$$

Hence,

$$\begin{aligned} 0 &= \int_{\mathbb{R}} \int_S \frac{1}{\sqrt{2\pi}\sigma} \int_{\mathbb{R}} h(x,y)^2 \exp(-(y-m)^2/2\sigma^2)\,\mathrm{d}y\,\mathrm{d}\phi(m,\sigma)\,\mathrm{d}\mu(x) \\ &= \int_{\mathbb{R}} \int_S q_h(x,m,\sigma)\,\mathrm{d}\phi(m,\sigma)\,\mathrm{d}\mu(x), \end{aligned}$$

implies that $q_h(x,m,\sigma) = 0$, for $\mu \otimes \phi$-a.e. $(x,m,\sigma) \in \mathbb{R} \times S$. As S has nonempty interior, $\mathrm{supp}(\mu)$ contains an open set, and q_h is a polynomial, this implies $q_h \equiv 0$. But then this in turn implies $h(x,y) = 0$ for all x, y, and therefore $h \equiv 0$. Hence the couple $(\boldsymbol{\lambda}, \boldsymbol{\phi})$ is a strictly feasible solution of (18), that is, Slater's condition[3] holds for (18). This in turn implies that there is not duality gap between (18) and its dual (19), i.e., $\tau_n = \tau_n^*$, and as $\tau_n \geq 0$, their value is finite. □

Theorem 9. *Let $S \subset \mathbb{R} \times \mathbb{R}_+$ as in* (17) *be compact, and let $\mu \in \mathscr{P}(\mathbb{R})$ be a probability measure such that* (4) *holds for some scalar $c > 0$.*

(i) *For every fixed n,* (18) *is a semidefinite relaxation of* (5) *(hence of* (3)*) and has an optimal solution $(\boldsymbol{\lambda}^{(n)}, \boldsymbol{\phi}^{(n)})$ with associated optimal value $\tau_n \leq \tau$ for all $n \geq n_0$.*

(ii) *For any accumulation point $(\boldsymbol{\lambda}^*, \boldsymbol{\phi}^*)$ of the sequence $(\boldsymbol{\lambda}^{(n)}, \boldsymbol{\phi}^{(n)})_{n\in\mathbb{N}}$ of optimal moment-sequences $(\boldsymbol{\lambda}^{(n)}, \boldsymbol{\phi}^{(n)})$ of* (18)*, $\boldsymbol{\lambda}^*$ (resp. $\boldsymbol{\phi}^*$) has a determinate representing measure λ^* on $\mathbb{R}^2$ (resp. ϕ^* on S) and (ϕ^*, λ^*) is an optimal solution of* (3) *and* (5). *That is:*

$$\tau_n \uparrow \tau = W_2(\mu - \nu_{\phi^*})^2 \quad \text{as } n \to \infty.$$

[3] Slater condition holds for the finite-dimensional conic program $\min_{\mathbf{x}}\{\mathbf{c}^T\mathbf{x} : \mathbf{A}\mathbf{x} = \mathbf{b};\, \mathbf{x} \in K\}$ for a linear mapping $A : \mathbb{R}^p \to \mathbb{R}^q$, vectors $\mathbf{c} \in \mathbb{R}^p$, $\mathbf{b} \in \mathbb{R}^q$, and a convex cone $K \subset \mathbb{R}^p$, if there exists an admissible solution $\mathbf{x}_0 \in \mathrm{int}(K)$.

For clarity of exposition a proof is postponed to Section 6.

Corollary 10. *Let τ_n and τ^*_n be as in* (18) *and* (19)*, respectively. Under the assumption in Theorem 9 and if S has nonempty interior and* $\mathrm{supp}(\mu)$ *contains an open set, then $\tau = \lim_{n\to\infty}\tau_n = \lim_{n\to\infty}\tau^*_n$, and therefore there is no duality gap between* (5) *and* (16)*, that is,*

$$\begin{aligned}&\inf_{\substack{\lambda\in\mathscr{P}(\mathbb{R}^2),\\ \phi\in\mathscr{P}(S)}}\left\{\int (x-y)^2\,\mathrm{d}\lambda : s.t. \quad \lambda_{j0}=\mu_j\,,\,\forall\, j\in\mathbb{N};\lambda_{0,j}-\int p_j(m,\sigma)\,\mathrm{d}\phi = 0\,,\,\forall\, j\in\mathbb{N}\right\}\\ &=\sup_{\substack{q\in\mathbb{R}[x],\\ g\in\mathbb{R}[y]}}\left\{\int q\,\mathrm{d}\mu : s.t. \quad q(x)+g(y)\leq (x-y)^2 \quad \forall x,y\in\mathbb{R};\sum_k g_k\,p_k(m,\sigma)\geq 0, \quad \forall (m,\sigma)\in S\right\}. \qquad (20)\end{aligned}$$

Proof. By Lemma (8), $\tau_n = \tau^*_n$ for all $n\geq n_0$, and by Theorem 9, $\tau = \lim_{n\to\infty}\tau_n$. Therefore τ^* in (16) is equal to τ, which yields (20). □

Observe that (20) resembles the usual duality in optimal transport when both marginals λ_x and λ_y are fixed; here the marginal λ_y is also part of the optimization via the mixing measure ϕ.

Corollary 11. *Let $S\subset\mathbb{R}\times\mathbb{R}_+$ be compact with nonempty interior and let $\mu\in\mathscr{P}(\mathbb{R})$ be such that* (4) *holds for some scalar $c>0$ and its support contains an open set. Then μ is a mixture of Gaussians with parameters $(m,\sigma)\in S$ if and only if for every $n\geq n_0$, $(q^*,g^*)=(0,0)$ and $\theta^*_j = 0$ for all $j=0\ldots,s$, is an optimal solution of* (19).

Proof. It μ is a Gaussian mixture with mixing measure $\phi^*\in\mathscr{P}(S)$, then $\tau = W_2(\mu,\nu_{\phi^*})=0$. As $0\leq\tau_n\leq\tau=0$ one obtains $\tau_n=\tau^*_n=0$ for all $n\geq n_0$ and $(q,g)=(0,0)$ with $\theta^*_j=0$ for all $j=0,\ldots,s$, is an obvious optimal solution of (19). □

3.3. *Recognizing a Gaussian mixture*

As a consequence of Corollary 11, if the input probability measure μ is *not* a mixture of Gaussian measures with parameters $(m,\sigma)\in S$, then the optimal value of (18) becomes positive at some step $n_*\geq n_0$ and then remains positive for all $n\geq n_*$ (as the sequence is monotone non decreasing). So the sequence of optimal values $(\tau_n)_{n\in\mathbb{N}}$ of the hierarchy (18) permits to detect in finitely many steps if μ is not a Gaussian mixture (with parameters $(m,\sigma)\in S$).

Recall that $d_j := \lceil\deg(u_j)/2\rceil$ and let $v:=\max_j d_j$.

Theorem 12. *With $S\subset\mathbb{R}^2$ as in* (17)*, let $\mu\in\mathscr{P}(\mathbb{R})$ be a given probability measure with finite moments $\boldsymbol{\mu}=(\mu_{(i,j)})_{(i;j)\in\mathbb{N}^2}$, and let τ and τ_n be as in* (3) *and* (18) *respectively.*

(i) *μ is a mixture of Gaussian measures, all with parameters $(m,\sigma)\in S$, if and only if $(\mu\otimes\mu,\phi^*)$ is an optimal solution of* (3) *for some $\phi^*\in\mathscr{P}(S)$. Moreover, $\tau_n=\tau=0$ for all $n\geq n_0$, i.e., the optimal value* 0 *is obtained at every step of the hierarchy of semidefinite relaxations* (18).

In addition, if μ is a mixture of finitely many (say r) Gaussian measures, all with parameters $(m,\sigma)\in S$, then for every n sufficiently large, the corresponding degree-$2n$ vector of moments $(\boldsymbol{\lambda}^*,\boldsymbol{\phi}^*)$ respectively associated with $\mu\otimes\mu$ and ϕ^*, is an optimal solution of (18) *and*

$$\mathrm{rank}(\mathbf{M}_n(\boldsymbol{\phi}^*)) = \mathrm{rank}(\mathbf{M}_{n-v}(\boldsymbol{\phi}^*)) = r\,. \qquad (21)$$

(ii) *Conversely, let $(\boldsymbol{\lambda}^*,\boldsymbol{\phi}^*)$ be an optimal solution of some degree-$2n$ relaxation* (18) *with $\tau_n=0$, and suppose that* (21) *holds. Then $\boldsymbol{\phi}^*$ is the degree-$2n$ moment vector of some*

r-atomic probability measure ϕ^* on S. Moreover, $\mu = \nu_{\phi^*}$ (i.e. μ is a Gaussian mixture with mixing measure ϕ^*) if and only if

$$\mu_j = \int p_j(m,\sigma)\,\mathrm{d}\phi^*\,, \quad \forall j > n+1\,. \tag{22}$$

Proof.

(i) ***Only if part.*** By definition there exists $\phi^* \in \mathscr{P}(S)$ such that

$$\mu(B) = \int_S \left(\frac{1}{\sqrt{2\pi}\sigma} \int_B \exp\left(\frac{-(x-m)^2}{2\sigma^2} \right) \mathrm{d}x \right) \mathrm{d}\phi^*(m,\sigma)\,, \quad \forall B \in \mathscr{B}(S)\,.$$

Then $\tau = W_2(\mu,\nu_{\phi^*}) = 0$, and with $\lambda^* := \mu \otimes \mu$, the couple (λ^*, ϕ^*) is an obvious optimal solution of (3). Moreover, $\tau_n = 0$ for all n, follows from $0 \leq \tau_n \leq \tau$ and $\tau = 0$.

If part. If $(\mu \otimes \mu, \phi^*)$ is an optimal solution of (3) then $\mu = \lambda_y = \nu_{\phi^*}$, i.e., μ is a Gaussian mixture with mixing measure $\phi^* \in \mathscr{P}(S)$, and $\tau = 0 = W_2(\mu,\nu_{\phi^*})^2$. Next, fix $n \geq n_0$ arbitrary. The finite vector of degree-$2n$ moments $(\boldsymbol{\lambda}^*, \boldsymbol{\phi}^*)$ of $\lambda^* = \mu \otimes \mu$ and ϕ^* respectively, is an obvious feasible solution of (18). Moreover $\lambda^*((x-y)^2) = \int (x-y)^2 \mathrm{d}\mu(x)\,\mathrm{d}\mu(y) = 0$, and as $\tau_n \geq 0$, $(\boldsymbol{\lambda}^*, \boldsymbol{\phi}^*)$ is an optimal solution of (18) with $\tau_n = 0$.

Next, as ϕ^* is r-atomic, $\operatorname{rank}(\mathbf{M}_n(\boldsymbol{\phi}^*)) = r$ for all sufficiently large n. As $\mathbf{M}_{n-v}(\boldsymbol{\phi}^*)$ is a submatrix of $\mathbf{M}_n(\boldsymbol{\phi}^*)$, (21) follows.

(ii). Conversely, if (22) holds at an optimal solution of a degree-$2n$ relaxation (18), then by Curto & Fialkow's flat extension theorem [15, Theorem 2.47], $\boldsymbol{\phi}^*$ is the degree-$2n$ moment sequence of some r-atomic $\phi^* \in \mathscr{P}(S)$. Next, $\tau_n = 0$ implies $\lambda^*((x-y)^2) = 0$ and so the vector $\mathbf{p} \in \mathbb{R}^{\binom{2+n}{2}}$ of coefficients of the polynomial $(x,y) \mapsto p(x,y) := (x-y) \in \mathbb{R}[x,y]_n$ is in the kernel of $\mathbf{M}_n(\boldsymbol{\lambda}^*)$ as

$$\langle \mathbf{p}, \mathbf{M}_n(\boldsymbol{\lambda}^*)\,\mathbf{p} \rangle = \lambda^*(p^2) = \lambda^*((x-y)^2) = 0\,.$$

That is, the second and third columns of $\mathbf{M}_n(\boldsymbol{\lambda}^*)$ (respectively indexed by the monomials x and y) are identical. In particular, this implies $\lambda^*_{(j,0)} = \lambda^*_{(0,j)}$ for all $j = 0,\dots n+1$. Equivalently $\mu_j = (\nu_{\phi^*})_j$ for all $j \leq n+1$, and therefore as μ is determinate, $\mu = \nu_{\phi^*}$ (and so $W_2(\mu,\nu_{\phi^*}) = 0$) if only if $\mu_j = \lambda^*_{(0,j)}$ for all j, and so if and only if (22) holds. □

The sufficient Curto & Fialkow's flatness condition (21) in Theorem 12 is very useful to detect whether μ is a Gaussian mixture ν_{ϕ^*} with an r-atomic mixing measure ϕ^* on S, in solving *finitely many* semidefinite relaxations. Indeed if (21) holds then it remains to check whether (22) holds (with no optimization involved).

Example 13. Let $S = [.07,1] \times [.02,1]$ and $\mu = r * \mathcal{N}(.1,.2) + (1-r) * \mathcal{N}(.5,.5)$ with $r \in (0,1)$. Then with $r = .2$ or $r = .3$, the atomic measure $\phi^* = r * \delta_{(.1,.2)} + (1-r) * \delta(.5,.5)$ is detected at step $n = 6$ of the semidefinite relaxation (18). Indeed, in its degree-12 optimal solution $(\boldsymbol{\lambda}^*, \boldsymbol{\phi}^*)$ obtained by running the GloptiPoly software [10] that implements the Moment-SOS hierarchy, $\boldsymbol{\phi}^*$ satisfies the flatness condition (21), and the atoms can be extracted by a linear algebra subroutine. However we could notice that if we enlarge the set S, then one needs to go to higher degrees in the hierarchy with potential numerical instabilities.

4. The multivariate case

The result in the univariate case extends to the multivariate case with μ on $\mathbb{R}^d$, provided that the set of parameters $(\mathbf{m},\Sigma) \in S \subset \mathbb{R}^d \times \mathbb{R}^{d(d+1)/2}$ is a compact basic semi-algebraic set. For instance one may consider the case where $(\mathbf{m},\Sigma) \in S$ with

$$S := \{(\mathbf{m},\Sigma) : a\mathbf{I} \preceq \Sigma \preceq b\mathbf{I}; \quad g_j(\mathbf{m}) \geq 0\,,\ j = 1,\dots,s\},$$

for some polynomials $g_j \in \mathbb{R}[m_1,\dots,m_d]$, $j=1,\dots,s$, and some given scalars $0<a<b$. Then using determinants of $\Sigma = (\sigma_{ij})_{i,j\leq d}$, the constraints $a\mathbf{I} \preceq \Sigma \preceq b\mathbf{I}$ reduces to $2d$ polynomials inequality constraints $q_k(\boldsymbol{\sigma}) \geq 0$, $k=1,\dots,2d$, with two of them of degree d. Then the set

$$S = \{(\mathbf{m},\boldsymbol{\sigma}) : g_j(\mathbf{m}) \geq 0,\ j=1,\dots,s;\ q_k(\boldsymbol{\sigma}) \geq 0,\ k=1,\dots,2d\} \subset \mathbb{R}^d \times \mathbb{R}^{d(d+1)/2}. \tag{23}$$

As S is compact and assuming one knows a scalar $R>0$ such that

$$R^2 - \|\mathbf{m}\|^2 - \|\boldsymbol{\sigma}\|^2 \geq 0, \quad \forall (\mathbf{m},\boldsymbol{\sigma}) \in S,$$

we may add the redundant quadratic constraint $R^2 - \|\mathbf{m}\|^2 - \|\boldsymbol{\sigma}\|^2 \geq 0$ (relabelled as $g_1(\mathbf{m},\boldsymbol{\sigma}) \geq 0$) in the definition (23) of S without changing S. In doing so, the quadratic module

$$Q(g,q) = \left\{ \sum_{j=0}^{s} \theta_j\, g_j + \sum_{k=1}^{2d} \theta'_k\, q_k : \theta_j, \theta'_k \in \Sigma[\mathbf{m},\boldsymbol{\sigma}] \right\} \tag{24}$$

is Archimedean. Next, as in the univariate case one introduces the polynomials $(p_{\boldsymbol{\alpha}} \in \mathbb{R}[\mathbf{m},\boldsymbol{\sigma}]_{|\boldsymbol{\alpha}|})_{\boldsymbol{\alpha}\in\mathbb{N}^d}$ defined by:

$$(\mathbf{m},\boldsymbol{\sigma}) \longmapsto p_{\boldsymbol{\alpha}}(\mathbf{m},\boldsymbol{\sigma}) := \frac{1}{(2\pi)^{d/2}\sqrt{\det(\Sigma)}} \int \mathbf{x}^{\boldsymbol{\alpha}} \exp(-(\mathbf{x}-\mathbf{m})^T \Sigma^{-1} (\mathbf{x}-\mathbf{m})/2)\, \mathrm{d}\mathbf{x}, \quad \boldsymbol{\alpha} \in \mathbb{N}^d, \tag{25}$$

as indeed every moment $\int \mathbf{x}^{\boldsymbol{\alpha}}\, \mathrm{d}\nu$ of a Gaussian probability measure $\nu = \mathcal{N}(\mathbf{m},\Sigma)$, is an explicit polynomial of its parameters $(\mathbf{m},\boldsymbol{\sigma})$, of total degree at most $|\alpha|$. Moreover, the marginal of a Gaussian measure $\mu = \mathcal{N}(\mathbf{m},\Sigma)$ with respect to x_i is the Gaussian measure $\mathcal{N}(m_i,\Sigma_{ii})$. Therefore

$$\int x_i^{2j}\, \mathrm{d}\mu = p_{2j}(m_i,\Sigma_{ii}), \quad \forall j \in \mathbb{N};\ i=1,\dots,d, \tag{26}$$

where p_{2j} has been defined in (10). Next, if $\mu \in \mathscr{P}(\mathbb{R}^d)$, the multivariate analogue of (3) reads:

$$\tau = \inf_{\phi\in\mathscr{P}(S)} W_2(\mu,\nu_\phi)^2 = \inf_{\phi\in\mathscr{P}(S),\lambda\in\mathscr{P}(\mathbb{R}^{2d})} \left\{ \int \|\mathbf{x}-\mathbf{y}\|^2\, \mathrm{d}\lambda : \lambda_{\mathbf{x}} = \mu;\ \lambda_{\mathbf{y}} = \nu_\phi \right\}. \tag{27}$$

and the analogue of the moment formulation (5) reads:

$$\inf_{\phi\in\mathscr{P}(S),\lambda\in\mathscr{P}(\mathbb{R}^{2d})} \{ \textstyle\int \|\mathbf{x}-\mathbf{y}\|^2\, \mathrm{d}\lambda : \lambda_{\boldsymbol{\alpha}0} = \mu_{\boldsymbol{\alpha}};\quad \lambda_{0\boldsymbol{\alpha}} - \phi(p_{\boldsymbol{\alpha}}(\mathbf{m},\boldsymbol{\sigma})) = 0, \quad \forall\, \boldsymbol{\alpha} \in \mathbb{N}^d \}. \tag{28}$$

Assumption 14.

(i) *The measure μ satisfies:* $\sup_i \int \exp(c\,|x_i|)\, \mathrm{d}\mu < \infty$ *for some* $c>0$.

(ii) *The set S in* (23) *is compact with nonempty interior, and the quadratic module* (24) *is Archimedean.*

Theorem 15. *Let Assumption 14 hold. Then:*

(i) *The optimal transport problem* (27) *has an optimal solution* $(\phi^*,\lambda^*) \in \mathscr{P}(S) \times \mathscr{P}(\mathbb{R}^{2d})$ *which is also an optimal solution of* (28). *Moreover both measures* $\lambda^* \in \mathscr{P}(\mathbb{R}^{2d})$ *and* $\nu_{\phi^*} \in \mathscr{P}(\mathbb{R}^d)$ *are moment determinate.*

(ii) *Moreover,* $\tau = 0$ *if and only if* $\lambda^* = \mu \otimes \mu$ *and* $\mu = \nu_{\phi^*}$, *i.e., μ is a Gaussian mixture with ϕ^* a mixing measure of parameters* $(\mathbf{m},\Sigma) \in S$.

Sketch of the proof. As in the proof of Theorem 5 in the univariate case let $(\lambda^{(n)},\phi^{(n)})_{n\in\mathbb{N}}$ be a minimizing sequence of (27). As S is compact there exists a subsequence $(n_k)_{k\in\mathbb{N}}$ and a probability measure $\phi^* \in \mathscr{P}(S)$ such that $\phi^{(n_k)} \Rightarrow \phi^*$ as $k\to\infty$.

Let $d' = d + d(d+1)/2$ and recall that $S \subset \mathbb{R}^{d'}$. Following exactly the same steps as in the proof of Theorem 5, there exists a subsequence denoted $(n'_\ell)_{\ell\in\mathbb{N}}$ and an infinite sequence $\boldsymbol{\lambda}^* = (\lambda^*_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{N}^{2d}}$, such that

$$\lim_{\ell\to\infty} \lambda_{\boldsymbol{\alpha}}^{(n'_\ell)} = \lambda^*_{\boldsymbol{\alpha}}, \quad \forall \boldsymbol{\alpha} \in \mathbb{N}^{2d}; \quad \lim_{\ell\to\infty} \phi_{\boldsymbol{\beta}}^{(n'_\ell)} = \phi^*_{\boldsymbol{\beta}}, \quad \forall \boldsymbol{\beta} \in \mathbb{N}^{d'}.$$

Moreover as S is compact and in view of (26), and (14)-(15), and by Corollary 4,

$$\sum_{j=1}^{\infty} \phi^*(p_{2j}(m_i,\Sigma_{ii}))^{-1/2j} = +\infty, \quad \forall i = 1,\ldots,d.$$

and therefore

$$\sum_{j=1}^{\infty} \lambda^*(y_i^{2j})^{-1/2j} = +\infty, \quad \forall i = 1,\ldots,d.$$

Next, by Assumption 14(i) on μ, one also has

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

and therefore the moment sequence $\boldsymbol{\lambda}^*$ satisfies multivariate Carleman's condition (see e.g. [15, Proposition 2.37]), which in turn implies that it is the moment sequence of some measure $\lambda^* \in \mathscr{M}(\mathbb{R}^{2d})_+$ which is moment determinate. Then again as in the proof of Theorem 5 we may conclude that (λ^*,ϕ^*) is an optimal solution of (15). □

Next, let $d_j = \lceil \deg(g_j)/2 \rceil$ and $t_k = \lceil \deg(q_k)/2 \rceil$, for all j and k. Then for every $n \geq n_0 = \max_{j,k}[d_j,\, t_k]$, the multivariate analogue of the semidefinite relaxation (18) reads:

$$\begin{aligned}\tau_n = \inf_{\boldsymbol{\phi},\boldsymbol{\lambda}} \Big\{ \int \|\mathbf{x}-\mathbf{y}\|^2\, \mathrm{d}\lambda : \lambda_{\boldsymbol{\alpha},0} = \mu_{\boldsymbol{\alpha}}; \quad \lambda_{0,\boldsymbol{\alpha}} - \phi(p_{\boldsymbol{\alpha}}(\mathbf{m},\boldsymbol{\sigma})) = 0, \quad \forall\, \boldsymbol{\alpha} \in \mathbb{N}^d_{2n}; \\ \mathbf{M}_n(\boldsymbol{\lambda}), \mathbf{M}_n(\boldsymbol{\phi}), \mathbf{M}_{n-d_j}(g_j \cdot \boldsymbol{\phi}), \mathbf{M}_{n-t_k}(q_k \cdot \boldsymbol{\phi}) \succeq 0; \quad j = 1,\ldots,s;\, k = 1,\ldots,2d \Big\}. \quad (29)\end{aligned}$$

Then an analogue of Theorem 9 holds and its proof is along the same lines. Also Curto & Fialkow's flatness condition [15, Theorem 2.47] is also valid in the multivariate setting. Similarly there is an exact analogue of Theorem 12.

5. Conclusion

We have considered Gaussian mixtures (with parameters (m,σ) in a given compact set S) closest in Wasserstein distance, to a given measure μ. Such Gaussian mixtures are optimal solutions of an infinite-dimensional optimal-transport linear program (LP) in which one marginal constraint contains the unknown mixing measure. Non-uniqueness is related to a classical identifiably issue. This LP can be solved by the Moment-SOS hierarchy, i.e., a sequence of semidefinite programs (convex relaxations) whose size increases with the number of moment constraints considered. That μ cannot be a Gaussian mixture is guaranteed to be detected at some step of the hierarchy. On the other hand if μ is a Gaussian mixture with mixing measure on S with finite support, a latter can sometimes be extracted from an optimal solution at some step of the hierarchy. In addition to the identifiability issue, an interesting research direction is concerned with whether a similar approach can be implemented when the distance is now measured in total variation instead of Wasserstein.

6. Appendix

In this paper we mainly use the $W_2(\mu,\nu)$-optimal transport problem (1) for two probability measures μ and ν, but we could also use the $W_1(\mu,\nu)$-optimal transport problem. Its primal formulation reads

$$W_1(\mu,\nu) = \inf_{\lambda \in \mathscr{P}(\mathbb{R}^2)} \left\{ \int_{\mathbb{R}^2} |x-y|\, \mathrm{d}\lambda(x,y) : \lambda_x = \mu;\, \lambda_y = \nu \right\},$$

while its dual formulation reads

$$W_1(\mu,\nu)=\sup_{f,g}\left\{\int_{\mathbb{R}^2} f(x)\,\mathrm{d}\mu(x)+\int g(y)\,\mathrm{d}\nu(y):\, f(x)+g(y)\leq |x-y|,\quad \forall x,y\in\mathbb{R}\right\}.$$

In order to proceed in a manner similar as for the W_2-distance, we need to write $\mathbb{R}^2=\overline{X_1\cup X_2}$ with $X_1:=\{(x,y):x<y\}$ and $X_2:=\{(x,y):x>y\}$, and impose $\lambda=\lambda_1+\lambda_2$ with $\mathrm{supp}(\lambda_1)=\overline{X_1}$ and $\mathrm{supp}(\lambda_2)=\overline{X_2}$.

6.1. *Proof of Theorem 5*

Proof.

(i). Let $(\lambda^{(n)},\phi^{(n)})_{n\in\mathbb{N}}\subset\mathscr{P}(\mathbb{R}^2)\times\mathscr{P}(S)$ be a minimizing sequence of (3) with $\rho_n:=W_2(\mu-\nu_{\phi^{(n)}})\downarrow\tau$ as n increases. As S is compact, the sequence $(\phi^{(n)})_{n\in\mathbb{N}}$ is tight and by Prohorov's theorem, there exists a subsequence $(n_k)_{k\in\mathbb{N}}$ and a probability measure $\phi^*\in\mathscr{P}(S)$ such that

$$\lim_{k\to\infty}\int h\,\mathrm{d}\phi^{(n_k)}=\int h\,\mathrm{d}\phi^*,\quad \forall h\in\mathscr{C}(S)\quad [\text{denoted } \phi_{n_k}\Rightarrow\phi^*].$$

In particular, $\phi^{(n_k)}_{(i,j)}\to\phi^*_{(i,j)}$ for all $(i,j)\in\mathbb{N}^2$. In addition, as $p_j\in\mathbb{R}[m,\sigma]$,

$$\lim_{k\to\infty}\lambda^{(n_k)}_{(0,j)}=\lim_{k\to\infty}\int p_j\,\mathrm{d}\phi^{(n_k)}=\int p_j\,\mathrm{d}\phi^*,\quad\forall j\in\mathbb{N},\tag{30}$$

and by feasibility, we also have

$$\lim_{k\to\infty}\lambda^{(n_k)}_{(j,0)}=\mu_j,\quad\forall j\in\mathbb{N}.$$

We want to prove that

$$\forall i,j\in\mathbb{N}:\quad \lim_{k\to\infty}\lambda^{(n_k)}_{(i,j)}=\int x^i y^i\,\mathrm{d}\lambda^*(x,y),$$

for some determinate measure λ^* on $\mathbb{R}^2$. That is, the vector of moments $\boldsymbol{\lambda}^{(n_k)}$ converges to the vector of moments of λ^*, and in particular

$$\lambda^*_{(j,0)}=\mu_j\quad\forall j\in\mathbb{N};\quad \lambda^*_{(0,j)}=\int_S p_j\,\mathrm{d}\phi^*=\int_{\mathbb{R}}x^j\,\mathrm{d}\nu_{\phi^*},\quad\forall j\in\mathbb{N}.$$

Notice that then (λ^*,ϕ^*) is an optimal solution of (3).

As S is compact, $|m|<M$ and $\sigma<M$ for some $M>0$ and therefore by (14),

$$\lambda^{(n)}_{(0,2j)}=\phi^{(n)}(p_{2j})\Longrightarrow\lambda^{(n)}_{(0,2j)}=\int_S p_{2j}\,\mathrm{d}\phi^{(n)}<(2Mj)^{2j}=:\rho_j,\quad\forall j\in\mathbb{N}.$$

This combined with $\lambda^{(n)}_{(2j,0)}=\mu_{2j}$ yields that the moment matrix $\mathbf{M}_k(\boldsymbol{\lambda}^{(n)})$ of the moment sequence $\boldsymbol{\lambda}^{(n)}$ of the measure $\lambda^{(n)}$ satisfies $\mathbf{M}_k(\boldsymbol{\lambda}^{(n)})\succeq 0$ for every k, and

$$\forall(k,\ell)\in\mathbb{N}^2\text{ with }k+\ell\leq 2j:\quad |\lambda^{(n)}_{k,\ell}|\leq\max[1,\mu_{2j},\rho_j]=:\rho'_j,\quad\forall j\in N.$$

See [13, Proposition 3.6, p. 60]. Then define the new infinite sequence $\widehat{\boldsymbol{\lambda}}^{(n)}$ by

$$\widehat{\lambda}^{(n)}_{(i,j)}:=\lambda_{(i,j)}/\rho'_k,\quad\forall(i,j)\text{ with }2k<i+j\leq 2k,\quad k=1,\ldots,\tag{31}$$

so that $\widehat{\boldsymbol{\lambda}}^{(n)}$ is an element of the unit ball of ℓ_∞, the Banach space of (uniformly) bounded sequences. As the unit ball $\mathbf{B}_{\ell_\infty}(0,1)$ of ℓ_∞ is sequentially compact in the weak-star topology $\sigma(\ell_\infty,\ell_1)$, there is a subsequence $(n'_\ell)_{\ell\in\mathbb{N}}\subset(n_k)_{k\in\mathbb{N}}$ and an infinite vector $\widehat{\boldsymbol{\lambda}}^*\in\mathbf{B}_{\ell_\infty}(0,1)$ such that (in particular)

$$\lim_{\ell\to\infty}\widehat{\lambda}^{(n'_\ell)}_{(i,j)}=\widehat{\lambda}^*_{(i,j)},\quad\forall(i,j)\in\mathbb{N}^2.$$

Then by the reverse scaling of (31) for $\widehat{\boldsymbol{\lambda}}^*$

$$\lim_{\ell\to\infty} \lambda^{(n'_\ell)}_{(i,j)} = \lambda^*_{(i,j)}; \quad \forall (i,j) \in \mathbb{N}^2, \tag{32}$$

for some infinite vector $\boldsymbol{\lambda}^* = (\lambda^*_{(i,j)})_{(i,j)\in\mathbb{N}^2}$. In addition, by (32), $\mathbf{M}_n(\boldsymbol{\lambda}^*) \succeq 0$ for all $n \in \mathbb{N}$, and

$$\lambda^*_{(j,0)} = \mu_j; \quad \lambda^*_{(0,j)} = \phi^*(p_j), \quad \forall j \in \mathbb{N},$$

and by Corollary 4,

$$\sum_{j=1}^{\infty} (\lambda^*_{(0,2j)})^{-1/2j} = +\infty.$$

As $\lambda^*_{(2j,0)} = \mu_{2j}$ for all $j \in \mathbb{N}$, and μ satisfies Carleman's condition, then by Theorem 1, $\boldsymbol{\lambda}^*$ has a representing measure λ^* on $\mathbb{R}^2$, which is moment determinate. This implies that (λ^*, ϕ^*) is a feasible solution of (3). Finally, as $(\lambda^{(n'_\ell)}, \phi^{(n'_\ell)})$ is a minimizing sequence of (3), then by (32),

$$\tau = \lim_{\ell\to\infty} \rho_{n_\ell} = \lim_{\ell\to\infty} \int (x-y)^2\, d\lambda^{(n'_\ell)} = \int (x-y)^2\, d\lambda^* \quad \text{[by (32)]},$$

which shows that (λ^*, ϕ^*) an optimal solution of (3).

Finally, in what precedes we have only used the respective moments $\boldsymbol{\lambda}^{(n)}$ and $\boldsymbol{\phi}^{(n)}$ of the measures $\lambda^{(n)}$ and $\phi^{(n)}$, and the constraints of (5). Hence by considering a minimizing sequence $(\lambda^{(n)}, \phi^{(n)})$ of (5) instead of (3), one reaches the same conclusion.

(ii) *If part.* Straightforward. Indeed if μ is a Gaussian mixture with ϕ^* a mixing measure of parameters $(m,\sigma) \in S$ then $\mu = \nu_{\phi^*}$ and with $\lambda^* := \mu \otimes \mu$, the couple (λ^*, ϕ^*) is a feasible solution of (3) with value $\tau = 0$, hence an optimal solution of (3).

Only if part. By (i) let (λ^*, ϕ^*) be an optimal solution of (3). As $0 = \tau = \int (x-y)^2\, d\lambda^*$, it follows that $\mathrm{supp}(\lambda^*) \subset \{(x,x) : x \in \mathbb{R}\}$, and therefore $\lambda^*_x = \lambda^*_y$, i.e., $\lambda^* = \mu \otimes \mu$, and therefore as $\lambda^*_y = \nu_{\phi^*}$, one obtains $\mu = \nu_{\phi^*}$, the desired result. □

6.2. *Proof of Theorem 9*

Proof.

(i). Let $(\boldsymbol{\lambda}, \boldsymbol{\phi})$ be a feasible solution of (18). As $g_1(m,\sigma) = R^2 - m^2 - \sigma^2$, the constraint $\mathbf{M}_{n-1}(g_1 \cdot \phi) \succeq 0$, implies that

$$\phi(\sigma^{2n}) < R^{2n}; \quad \phi(m^{2n}) < R^{2n}.$$

By [13, Proposition 3.6, p. 60], this combined with $\mathbf{M}_n(\boldsymbol{\phi}) \succeq 0$, and $\phi_{(0,0)} = 1$, yields $|\phi_{(i,j)}| < \max[1, R^{2n}]$ for all i,j with $i+j \le 2n$. Moreover, $\lambda_{2n,0} = \mu_{2n}$, and by (14),

$$\lambda_{(0,2n)} = \phi(p_{2n}) < (2nR)^{2n} =: \rho_n.$$

Again by [13, Proposition 3.6, p. 60], for all (i,j) with $i+j \le 2n$,

$$|\lambda_{(i,j)}| < \max[1, \lambda_{(2n,0)}, \lambda_{(0,2n)}] < \max[1, \mu_{2n}, \rho_n] =: \rho'_n,$$

which implies that the feasible set of (18) in compact, and therefore (18) has an optimal solution $(\boldsymbol{\lambda}^{(n)}, \boldsymbol{\phi}^{(n)})$ with value $\lambda^{(n)}((x-y)^2) = \tau_n$, and as (18) is a relaxation of (5), $0 \le \tau_n \le \tau$ for all n.

(ii). Complete the finite vector $\boldsymbol{\lambda}^{(n)}$ (resp. $\boldsymbol{\phi}^{(n)}$) with zeros to make it an infinite sequence $\boldsymbol{\lambda}^{(n)} = (\lambda^{(n)}_{(i,j)})_{(i,j)\in\mathbb{N}^2}$ (resp. $\boldsymbol{\phi}^{(n)} = (\boldsymbol{\phi}^{(n)}_{(i,j)})_{(i,j)\in\mathbb{N}^2}$). Then define the new infinite sequences $\widehat{\boldsymbol{\lambda}}^{(n)}$ and $\widehat{\boldsymbol{\phi}}^{(n)}$ by

$$\begin{aligned} \widehat{\lambda}^{(n)}_{(i,j)} &:= \lambda_{(i,j)}/\rho_k, \quad \forall (i,j) \text{ with } 2k < i+j \le 2k, \quad k = 1, \ldots \\ \widehat{\phi}^{(n)}_{(i,j)} &:= \phi_{(i,j)}/R^{2k}, \quad \forall (i,j) \text{ with } 2k < i+j \le 2k, \quad k = 1, \ldots, \end{aligned} \tag{33}$$

so that $\widehat{\boldsymbol{\lambda}}^{(n)}$ is an element of the unit ball of ℓ_∞, the Banach space of (uniformly) bounded sequences, and smililarly for $\widehat{\boldsymbol{\phi}}^{(n)}$. Again, as the unit ball of ℓ_∞ is sequentially compact in the weak-star topology $\sigma(\ell_\infty, \ell_1)$, there is a subsequence $(n_k)_{k\in\mathbb{N}}$ and infinite vectors $\widehat{\boldsymbol{\lambda}}^* \in \ell_\infty$ and $\widehat{\boldsymbol{\phi}}^* \in \ell_\infty$ such that

$$\lim_{k\to\infty} \widehat{\lambda}^{(n_k)}_{(i,j)} = \widehat{\lambda}^*_{(i,j)}; \quad \lim_{k\to\infty} \widehat{\phi}^{(n_k)}_{(i,j)} = \widehat{\phi}^*_{(i,j)}, \quad \forall (i,j) \in \mathbb{N}^2.$$

Then by the reverse scaling of (33) for $\widehat{\boldsymbol{\lambda}}^*$ and $\widehat{\boldsymbol{\phi}}^*$,

$$\lim_{k\to\infty} \lambda^{(n_k)}_{(i,j)} = \lambda^*_{(i,j)}; \quad \lim_{k\to\infty} \phi^{(n_k)}_{(i,j)} = \phi^*_{(i,j)}, \quad \forall (i,j) \in \mathbb{N}^2, \tag{34}$$

for some infinite vectors $\boldsymbol{\lambda}^*$ and $\boldsymbol{\phi}^*$. Next, by (34), $\mathbf{M}_n(\boldsymbol{\lambda}^*) \succeq 0$, $\mathbf{M}_n(\boldsymbol{\phi}^*) \succeq 0$, and $\mathbf{M}_n(g_j \cdot \boldsymbol{\phi}^*) \succeq 0$ for all n, with

$$\lambda^*_{(j,0)} = \mu_j \quad \text{and} \quad \lambda^*_{(0,j)} = \phi^*(p_j), \quad \forall j \in \mathbb{N}.$$

As $g_1(m,\sigma) = R^2 - m^2 - \sigma^2$, the quadratic module

$$Q(g) = \left\{ \sum_{j=0}^{s} \theta_j(m,\sigma)\, g_j(m,\sigma) : \theta_j \in \Sigma[m,\sigma] \right\}$$

is Archimedean and therefore, by Putinar's Positivstellensatz [24], $\boldsymbol{\phi}^*$ has a representing measure on S. Moreover, as in the proof of Theorem 5, by Corollary 4,

$$\sum_{j=1}^{\infty} (\lambda^*_{(0,2j)})^{-1/2j} = +\infty,$$

and as $\lambda^*_{(2j,0)} = \mu_{2j}$ for all $j \in \mathbb{N}$, and μ satisfies Carleman's condition, then by Theorem 1, $\boldsymbol{\lambda}^*$ has a representing measure λ^* on $\mathbb{R}^2$, which is moment determinate. In particular its marginal λ^*_y with moments $(\lambda^*_{(0,j)})_{j\in\mathbb{N}}$ is also moment determinate. Next, let ν_{ϕ^*} be the measure on $\mathbb{R}$ with Gaussian mixture ϕ^*. As

$$\lambda^*_{(0,j)} = \phi^*(p_j) = \int x^j \, \mathrm{d}\nu_{\phi^*}(x), \quad \forall j \in \mathbb{N},$$

and as λ^*_y is moment determinate, this show that $\lambda^*_y = \nu_{\phi^*}$. Hence (ϕ^*, λ^*) is feasible for (3) with value $\lambda^*((x-y)^2)$. In addition, as $\tau_n \leq \tau$ for all n,

$$\tau \leq \lambda^*(x-y)^2 = \lim_{\ell\to\infty} \lambda^{(n'_\ell)}((x-y)^2) \quad ([\text{by (34)}]) = \lim_{\ell\to\infty} \tau_{n'_\ell} \leq \tau,$$

so that $\tau = \lambda^*((x-y)^2)$, and therefore (λ^*, ϕ^*) is an optimal solution of (3) (and of (5) as well). □

Declaration of interests

The authors do not work for, advise, own shares in, or receive funds from any organization that could benefit from this article, and have declared no affiliations other than their research organizations.

References

[1] *Handbook on semidefinite, conic and polynomial optimization*, (M. F. Anjos and J. B. Lasserre, eds.), Springer, 2012, pp. xii+960.

[2] C. Améndola, J.-C. Faugère and B. Sturmfels, "Moment varieties of Gaussian mixtures", *J. Algebr. Stat.* **7** (2016), no. 1, pp. 14–28.

[3] C. Améndola, K. Ranestad and B. Sturmfels, "Algebraic identifiability of Gaussian mixtures", *Int. Math. Res. Not.* **2018** (2018), no. 21, pp. 6556–6580.

[4] A. Bakshi, I. Diakonikolas, H. Jia, D. M. Kane, P. K. Kothari and S. S. Vempala, “Robustly learning mixtures of k arbitrary Gaussians”, in *STOC ’22—Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, Association for Computing Machinery, 2022, pp. 1234–1247.

[5] X. Bin, F. Bunea and J. Niles-Wred, “Estimation and inference for the Wasserstein distance between mixing measures in topic models”, 2023. https://arxiv.org/abs/2206.12768.

[6] M. Di Zio, U. Guarnera and R. Rocci, “A mixture of mixture models for a classification problem: the unity measure error”, *Comput. Stat. Data Anal.* **51** (2007), no. 5, pp. 2573–2585.

[7] C. F. Dunkl and Y. Xu, *Orthogonal polynomials of several variables*, Cambridge University Press, 2001, pp. xvi+390.

[8] P. Heinrich and J. Kahn, “Strong identifiability and optimal minimax rates for finite mixture estimation”, *Ann. Stat.* **46** (2018), no. 6A, pp. 2844–2870.

[9] D. Henrion, M. Korda and J. B. Lasserre, *The moment-SOS hierarchy—lectures in probability, statistics, computational geometry, control and nonlinear PDEs*, World Scientific, 2021, pp. xvii+229.

[10] D. Henrion, J. B. Lasserre and J. Löfberg, “GloptiPoly 3: moments, optimization and semidefinite programming”, *Optim. Methods Softw.* **24** (2009), no. 4-5, pp. 761–779.

[11] A. T. Kalai, A. Moitra and G. Valiant, “Efficiently learning mixtures of two Gaussians”, in *STOC’10—Proceedings of the 2010 ACM International Symposium on Theory of Computing*, Association for Computing Machinery, 2010, pp. 553–562.

[12] P. K. Kothari, P. Manohar and B. H. Zhang, “Polynomial-time sum-of-squares can robustly estimate mean and covariance of Gaussians optimally”, in *Proceedings of The 33rd International Conference on Algorithmic Learning Theory* (S. Dasgupta and N. Haghtalab, eds.), PMLR, 2022, pp. 638–667.

[13] J. B. Lasserre, *Moments, positive polynomials and their applications*, Imperial College Press, 2010, pp. xxii+361.

[14] J. B. Lasserre, “The existence of Gaussian cubature formulas”, *J. Approx. Theory* **164** (2012), no. 5, pp. 572–585.

[15] J. B. Lasserre, *An introduction to polynomial and semi-algebraic optimization*, Cambridge University Press, 2015, pp. xiv+339.

[16] B. G. Lindsay, “Moment matrices: applications in mixtures”, *Ann. Stat.* **17** (1989), no. 2, pp. 722–740.

[17] J. S. Marron and M. P. Wand, “Exact mean integrated squared error”, *Ann. Stat.* **20** (1992), no. 2, pp. 712–736.

[18] G. McLachlan and D. Peel, *Finite mixture models*, John Wiley & Sons, 2000.

[19] A. Moitra and G. Valiant, “Settling the polynomial learnability of mixtures of Gaussians”, in *2010 IEEE 51st Annual Symposium on Foundations of Computer Science—FOCS 2010*, IEEE Computer Society, 2010, pp. 93–102.

[20] H. M. Möller, “On square positive extensions and cubature formulas”, *J. Comput. Appl. Math.* **199** (2007), no. 1, pp. 80–88.

[21] X. Nguyen, “Convergence of latent mixing measures in finite and infinite mixture models”, *Ann. Stat.* **41** (2013), no. 1, pp. 370–400.

[22] R. O’Donnell, “SOS is not obviously automatizable, even approximately”, in *8th Innovations in Theoretical Computer Science Conference*, Schloss Dagstuhl. Leibniz-Zent. Inform., 2017. No. 59, 10 pages.

[23] H. Permuter, J. Francos and I. Jermyn, “A study of Gaussian mixture models of color and texture features for image classification and segmentation”, *Pattern Recognition* **39** (2006), no. 4, pp. 695–706.

[24] M. Putinar, “Positive polynomials on compact semi-algebraic sets”, *Indiana Univ. Math. J.* **42** (1993), no. 3, pp. 969–984.

[25] J. Wang, “Generating daily changes in market variables using a multivariate mixture of normal distributions”, in *Proceeding of the 2001 Winter Simulation Conference (Cat. No.01CH37304)*, 2001, pp. 283–289.

[26] Y. Wu and P. Yang, “Optimal estimation of Gaussian mixtures via denoised method of moments”, *Ann. Stat.* **48** (2020), no. 4, pp. 1981–2007.

[27] G. Yu, G. Sapiro and S. Mallat, “Solving inverse problems with piecewise linear estimators: from Gaussian mixture models to structured sparsity”, *IEEE Trans. Image Process.* **21** (2012), no. 5, pp. 2481–2499.