Mean squared error minimization
for inverse moment problems[*]

Didier Henrion[1,2,3], Jean B. Lasserre[1,2,4], Martin Mevissen[5]

September 3, 2012

Abstract

We consider the problem of approximating the unknown density $u \in L^2(\Omega, \lambda)$ of a measure μ on $\Omega \subset \mathbb{R}^n$, absolutely continuous with respect to some given reference measure λ, from the only knowledge of finitely many moments of μ. Given $d \in \mathbb{N}$ and moments of order d, we provide a polynomial p_d which minimizes the mean square error $\int (u - p)^2 d\lambda$ over all polynomials p of degree at most d. If there is no additional requirement, p_d is obtained as solution of a linear system. In addition, if p_d is expressed in the basis of polynomials that are orthonormal with respect to λ, its vector of coefficients is just the vector of given moments and no computation is needed. Moreover $p_d \to u$ in $L^2(\Omega, \lambda)$ as $d \to \infty$. In general nonnegativity of p_d is not guaranteed even though u is nonnegative. However, with this additional nonnegativity requirement one obtains analogous results but computing $p_d \geq 0$ that minimizes $\int (u - p)^2 d\lambda$ now requires solving an appropriate semidefinite program. We have tested the approach on some applications arising from the reconstruction of geometrical objects and the approximation of solutions of nonlinear differential equations. In all cases our results are significantly better than those obtained with the maximum entropy technique for estimating u.

Keywords: Moment problems; density estimation; inverse problems; semidefinite programming.

[1]CNRS, LAAS, 7 avenue du colonel Roche, F-31400 Toulouse, France.

[2]Univ. de Toulouse, LAAS, F-31400 Toulouse, France.

[3]Faculty of Electrical Engineering, Czech Technical University in Prague, Technická 2, CZ-16626 Prague, Czech Republic.

[4]Institut de Mathématiques de Toulouse, Univ. de Toulouse, UPS, F-31400 Toulouse, France.

[5]IBM Research - Ireland, Dublin Technology Campus, Damastown Ind. Park, Mulhuddart, Dublin 15, Ireland.

1 Introduction

Estimating the density u of an unknown measure μ is a well-known problem in statistical analysis, physics or engineering. In a statistical context, one is usually given observations in the form of a sample of independent or dependent identically distributed random variables obtained from the unknown measure μ. And so there has been extensive research on estimating the density based on these observations. For instance, in one of the most popular approaches, the kernel density estimation [25], the density u is estimated via a linear combination of kernel functions - each of them being identified with exactly one observation. The crucial step in this method is to choose an appropriate bandwidth for which minimizing the integrated or the mean-integrated squared error between u and its estimate is most common. Another very popular approach uses wavelets [16, 6, 30], an example of approximating a density by a truncated orthonormal series. The coefficients in the truncated wavelet expansion are moment estimates derived from the given identically distributed observations. Again, the approximation accuracy is often measured by the mean-integrated squared error and depends on the number of observations and the degree of the truncation. This approach provides a global density estimate satisfying both good local and periodic approximation properties. For further details the interested reader is referred to [7, 11, 30] and the many references therein.

In another context - arising in challenging fields such as image recognition, solving nonlinear differential equations, spectral estimation or speech processing - no direct observation is available, but rather finitely many moments of the unknown measure μ are given. Then the issue is to reconstruct or approximate the density u based on the only knowledge of finitely many moments, (say up to order $d \in \mathbb{N}$), an *inverse* problem from moments. A simple method due to [29] approximates the density u by a polynomial p of degree at most d, so that the moments of the measure $pd\lambda$ matches those of μ, up to order d. However, and in contrast with more sophisticated approaches, the resulting polynomial approximation p is not guaranteed to be a density (even though u is) as it may takes negative values on the domain of integration. One classical approach to the moment problem is the Padé approximation [3] which is based on approximating the measure by a (finite) linear combination of Dirac measures. The Dirac measures and their weights in the decomposition are determined by solving a nonlinear system of equations. In the maximum entropy estimation (another classical approach) one selects the best approximation of u by maximizing some functional entropy, the most popular being the Boltzmann-Shannon entropy. In general some type of weak convergence takes place as the degree increases as detailed in [5]. Alternatively the norm of the approximate density is chosen as an objective function [4, 28, 15, 9], which allows to show a stronger convergence in norm. In [21], maximum entropy and Padé approximates have been compared on some numerical experiments. Finally, piecewise polynomial spline based approaches have also been proposed in [14].

Motivation. Our main motivation to study the (inverse) moment problem arises in the context of the so-called generalized problem of moments (GPM). The abstract GPM is a infinite-dimensional linear program on some space of Borel measures on $\mathbb{R}^n$ and its applications seem endless, see e.g. [17, 18] and the many references therein. For instance, to cite a few applications, the GPM framework can be used to help solve a weak

formulation of some ordinary or partial differential equations, as well as some calculus of variations and optimal control problems. The solution u of the original problem (or an appropriate translate) is interpreted as a density with respect to the Lebesgue measure λ on some domain and one computes (or approximates) finitely many moments of the measure $d\mu := u d\lambda$ by solving an appropriate finite-dimensional optimization problem. But then to recover an approximate solution of the original problem one has to solve an inverse problem from moments. This approach is particularly attractive when the data of the original problem consist of polynomials and basic semi-algebraic sets. In this case one may define a hierarchy (as the number of moments increases) of so-called semidefinite programs to compute approximations of increasing quality.

Contribution

In this paper we consider the following inverse problem from moments: Let μ be a finite Borel measure absolutely continuous with respect to some reference measure λ on a box Ω of $\mathbb{R}^n$ and whose density u is assumed to be in $L^2(\Omega, \lambda)$, with no continuity assumption as in previous works. The ultimate goal is to compute an *approximation* u_d of u, based on the only knowledge of finitely many moments (say up to order d) of μ. In addition, for consistency, it would be highly desirable to also obtain some "convergence" $u_d \to u$ as $d \to \infty$.

(a) Firstly, we approximate the density u by a polynomial u_d^* of degree d which minimizes the mean squared error $\int_\Omega (u - p)^2 d\lambda$ (or equivalently the $L^2(\Omega, \lambda)$-norm $\|u - p\|_2^2$) over all polynomials p of degree at most d. We show that an unconstrained L^2-norm minimizer u_d^* exists, is unique, and coincides with the simple polynomial approximation due to [29]; it can be determined by solving a system of linear equations. It turns out that u_d^* matches all moments up to degree d, and it is even easier to compute if it is expressed in the basis of polynomials that are orthonormal with respect to λ. No inversion is needed and the coefficients of u_d^* in such a basis are just the given moments. Moreover we show that $u_d^* \to u$ in $L^2(\Omega, \lambda)$ as $d \to \infty$, which is the best we can hope for in general since there is no continuity assumption on u; in particular $u_{d_k}^* \to u$ almost-everywhere and almost-uniformly on Ω for some subsequence (d_k), $k \in \mathbb{N}$. Even though both proofs are rather straightforward, to the best of our knowledge it has not been pointed out before that not only this mean squared error estimate u_d^* is much easier to compute than the corresponding maximum entropy estimate, but it also converges to u as $d \to \infty$ in a much stronger sense. For the *univariate case*, in references [27] and [2] the authors address the problem of approximating a *continuous* density on a compact interval by polynomials or kernel density functions that match a fixed number of moments. In this case, convergence in supremum norm is obtained when the number of moments increases. An extension to the noncompact (Stieltjes) case is carried out in [8]. Notice that in [27] it was already observed that the resulting polynomial approximation also minimizes the mean square error and its coefficients solve a linear system of equations. In [4, 28] the minimum-norm solution (and not the minimum distance solution) is shown to be unique solution of a system of linear equations. In [15] the minimal distance solution is considered but it is obtained as the solution of a constrained optimization problem and requires an initial guess for the density estimate.

3

(b) However, as already mentioned and unlike the maximum entropy estimate, the above unconstrained L^2-norm minimizer u_d^* may *not* be a density as it may take negative values on Ω. Of course, the nonnegative function $u_{d+}^* := \max[0, u_d^*]$ also converges to u in L^2 but it is not a polynomial anymore. So we next propose to obtain a *nonnegative polynomial* approximation u_d^* by minimizing the same L^2-norm criterion but now under the additional constraint that the candidate polynomial approximations should be nonnegative on Ω. In principle such a constraint is difficult to handle which probably explains why it has been ignored in previous works. Fortunately, if Ω is a compact basic semi-algebraic set one is able to enforce this positivity constraint by using Putinar's Positivstellensatz [26] which provides a nice positivity certificate for polynomials strictly positive on Ω. Importantly, the resulting optimization problem is convex and even more, a semidefinite programming (SDP) problem which can be solved efficiently by public domain solvers based on interior-point algorithms. Moreover, again as in the unconstrained case, we prove the convergence $u_d^* \to u$ in $L^2(\Omega, \lambda)$ as $d \to \infty$ (and so almost-everywhere and almost-uniform convergence on Ω as well for some subsequence $(u_{d_k}^*)$, $k \in \mathbb{N}$) which is far stronger than the weak convergence obtained for the maximum entropy estimate. Notice, in [29, 27] methods for obtaining some non-negative estimates are discussed, however these estimates do not satisfy the same properties in terms of mean-square error minimization and convergence as in the unconstrained case. In the kernel density element method [2, 8] a nonnegative density estimate for the univariate case is obtained by solving a constrained convex quadratic optimization problem. However, requiring each coefficient in the representation to be nonnegative as presented there seems more restrictive than the nonnegative polynomial approximation proposed in this paper.

(c) Our approach is illustrated on some challenging applications. In the first set of problems we are concerned with recovering the shape of geometrical objects whereas in the second set of problems we approximate solutions of nonlinear differential equations. Moreover, we demonstrate the potential of this approach for approximating densities with jump discontinuities, which is harder to achieve than for the smooth, univariate functions discussed in [27]. The resulting L^2-approximations clearly outperform the maximum entropy estimates with respect to both running time and pointwise approximation accuracy. Moreover, our approach is able to handle sets Ω more complicated than a box (as long as the moments of the measure $u d\lambda$ are available) as support for the unknown density, whereas such sets are a challenge for computing maximum entropy estimates because integrals of a combination of polynomials and exponentials of polynomials must be computed repeatedly.

Outline of the paper

In Section 2 we introduce the notation and we state the problem to be solved. In Section 3 we present our approach to approximate an unknown density u by a polynomial u_d^* of degree at most d via unconstrained and constrained L^2-norm minimization, respectively; in both cases we also prove the convergence $u_d^* \to u$ in $L^2(\Omega, \lambda)$ (and almost-uniform convergence on Ω as well for some subsequence) as d increases. In Section 4 we illustrate the approach on a number of examples - most notably from recovering geometric objects and approximating solutions of nonlinear differential equations - and highlight its advantages

4

when compared with the maximum entropy estimation. Finally, we discuss methods to improve the stability of our approach by considering orthogonal bases for the functions spaces we use to approximate the density. And we discuss the limits of approximating discontinuous functions by smooth functions in connection with the well-known Gibbs effect.

2 Notation, definitions and preliminaries

2.1 Notation and definitions

Let $\mathbb{R}[\mathbf{x}]$ (resp. $\mathbb{R}[\mathbf{x}]_d$) denote the ring of real polynomials in the variables $\mathbf{x} = (x_1, \ldots, x_n)$ (resp. polynomials of degree at most d), whereas $\Sigma[\mathbf{x}]$ (resp. $\Sigma[\mathbf{x}]_d$) denotes its subset of sums of squares (SOS) polynomials (resp. SOS of degree at most $2d$).

With $\Omega \subset \mathbb{R}^n$ and a given reference measure λ on Ω, let $L^2(\Omega, \lambda)$ be the space of functions on Ω whose square is λ-integrable and let $L^2_+(\Omega, \lambda) \subset L^2(\Omega, \lambda)$ be the convex cone of nonnegative elements. Let $C(\Omega)$ (resp. $C_+(\Omega)$) be the space of continuous functions (resp. continuous nonnegative functions) on Ω. Let $P(\Omega)$ be the space of polynomials nonnegative on Ω.

For every $\alpha \in \mathbb{N}^n$ the notation $\mathbf{x}^\alpha$ stands for the monomial $x_1^{\alpha_1} \cdots x_n^{\alpha_n}$ and for every $d \in \mathbb{N}$, let $\mathbb{N}^n_d := \{\alpha \in \mathbb{N}^n : \sum_j \alpha_j \leq d\}$ whose cardinal is $s(d) = \binom{n+d}{d}$. A polynomial $f \in \mathbb{R}[\mathbf{x}]$ is written

$$\mathbf{x} \mapsto f(\mathbf{x}) = \sum_{\alpha \in \mathbb{N}^n} f_\alpha \mathbf{x}^\alpha$$

and f can be identified with its vector of coefficients $\mathbf{f} = (f_\alpha)$ in the canonical basis $(\mathbf{x}^\alpha)$, $\alpha \in \mathbb{N}^n$. Denote by $\mathbb{S}^n$ the space of real $n \times n$ symmetric matrices, and by $\mathbb{S}^n_+$ the cone of positive semidefinite elements of $\mathbb{S}^n$. For any $\mathbf{A} \in \mathbb{S}^n_+$ the notation $\mathbf{A} \succeq 0$ stands for positive semidefinite. A real sequence $\mathbf{y} = (y_\alpha)$, $\alpha \in \mathbb{N}^n$, has a representing measure if there exists some finite Borel measure μ on $\mathbb{R}^n$ such that

$$y_\alpha = \int \mathbf{x}^\alpha \, d\mu(\mathbf{x}), \qquad \forall \alpha \in \mathbb{N}^n.$$

Linear functional

Given a real sequence $\mathbf{y} = (y_\alpha)$ define the Riesz linear functional $L_\mathbf{y} : \mathbb{R}[\mathbf{x}] \to \mathbb{R}$ by:

$$f \left(= \sum_\alpha f_\alpha \mathbf{x}^\alpha \right) \quad \mapsto L_\mathbf{y}(f) = \sum_\alpha f_\alpha \, y_\alpha, \qquad f \in \mathbb{R}[\mathbf{x}].$$

Moment matrix

Given $d \in \mathbb{N}$, the moment matrix of order d associated with a sequence $\mathbf{y} = (y_\alpha)$, $\alpha \in \mathbb{N}^n$, is the real symmetric matrix $\mathbf{M}_d(\mathbf{y})$ with rows and columns indexed by $\mathbb{N}^n_d$, and whose

5

entry (α, β) is $y_{\alpha+\beta}$, for every $\alpha, \beta \in \mathbb{N}_d^n$. If $\mathbf{y}$ has a representing measure μ then $\mathbf{M}_d(\mathbf{y}) \succeq 0$ because

$$\langle \mathbf{f}, \mathbf{M}_d(\mathbf{y})\mathbf{f} \rangle \; = \; \int f^2 \, d\mu \geq 0, \qquad \forall \mathbf{f} \in \mathbb{R}^{s(d)}.$$

Localizing matrix

With $\mathbf{y}$ as above and $g \in \mathbb{R}[\mathbf{x}]$ (with $g(\mathbf{x}) = \sum_\gamma g_\gamma \mathbf{x}^\gamma$), the localizing matrix of order d associated with $\mathbf{y}$ and g is the real symmetric matrix $\mathbf{M}_d(g\,\mathbf{y})$ with rows and columns indexed by $\mathbb{N}_d^n$, and whose entry (α, β) is $\sum_\gamma g_\gamma y_{(\alpha+\beta+\gamma)}$, for every $\alpha, \beta \in \mathbb{N}_d^n$. If $\mathbf{y}$ has a representing measure μ whose support is contained in the set $\{\mathbf{x} \, : \, g(\mathbf{x}) \geq 0\}$ then $\mathbf{M}_d(g\,\mathbf{y}) \succeq 0$ because

$$\langle \mathbf{f}, \mathbf{M}_d(g\,\mathbf{y})\mathbf{f} \rangle \; = \; \int f^2 \, g \, d\mu \geq 0, \qquad \forall \mathbf{f} \in \mathbb{R}^{s(d)}.$$

2.2   Problem statement

We consider the following setting. For $\Omega \subset \mathbb{R}^n$ compact, let μ and λ be σ-finite Borel measures supported on Ω. Assume that the moments of λ are known and μ is absolutely continuous with respect to λ ($\mu \ll \lambda$) with Radon-Nikodým derivative (or density) $u : \Omega \to \mathbb{R}_+$, with respect to λ. The density u is unknown but we know finitely many moments $\mathbf{y} = (y_\alpha)$ of μ, that is,

$$y_\alpha := \int_\Omega \mathbf{x}^\alpha u(\mathbf{x}) d\lambda(\mathbf{x}) = \int_\Omega \mathbf{x}^\alpha d\mu(\mathbf{x}), \qquad \forall \alpha \in \mathbb{N}_d^n, \tag{1}$$

for some $d \in \mathbb{N}$.

The issue is to find an estimate $u_d : \Omega \to \mathbb{R}_+$ for u, such that

$$\int_\Omega \mathbf{x}^\alpha u_d(\mathbf{x}) d\lambda(\mathbf{x}) = y_\alpha, \qquad \forall \alpha \in \mathbb{N}_d^n. \tag{2}$$

2.3   Maximum entropy estimation

We briefly describe the maximum entropy method due to [12, 13, 5] as a reference for later comparison with the mean squared error approach.

If one chooses the Boltzmann-Shannon entropy $H(u) := -u \log u$, the resulting estimate with maximum-entropy is an optimal solution of the optimization problem

$$\max_{u_d} \int_\Omega H(u_d) d\lambda \text{ s.t. } \int_\Omega \mathbf{x}^\alpha u_d(\mathbf{x}) d\lambda(\mathbf{x}) = y_\alpha, \quad \forall \alpha \in \mathbb{N}_d^n.$$

It turns out that an optimal solution u_d^* is of the form

$$\mathbf{x} \mapsto u_d^*(\mathbf{x}) := \exp\left(\sum_{|\alpha| \leq d} u_\alpha \mathbf{x}^\alpha\right)$$

for some vector $\mathbf{u}_d = (u_\alpha) \in \mathbb{R}^{s(d)}$. Hence, computing an optimal solution u_d^* reduces to solving the finite-dimensional convex optimization problem

$$\max_{\mathbf{u}_d \in \mathbb{R}^{s(d)}} \left\{ \langle \mathbf{y}, \mathbf{u}_d \rangle - \int_\Omega \exp\left(\sum_{|\alpha| \leq d} u_\alpha \mathbf{x}^\alpha \right) d\lambda(\mathbf{x}) \right\} \tag{3}$$

where $\mathbf{y} = (y_\alpha)$ is the given moment information on the unknown density u. If (u_d^*), $d \in \mathbb{N}$, is a sequence of optimal solutions to (3), then following weak convergence occurs:

$$\lim_{d \to \infty} \int_\Omega \psi(\mathbf{x}) u_d^*(\mathbf{x}) \, d\lambda(\mathbf{x}) = \int_\Omega \psi(\mathbf{x}) \, u(\mathbf{x}) \, d\lambda(\mathbf{x}), \tag{4}$$

for all bounded measurable functions $\psi : \Omega \to \mathbb{R}$ continuous almost everywhere. For more details the interested reader is referred to [5].

Since the estimate u_d^* is an exponential of a polynomial, it is guaranteed to be nonnegative on Ω and so it is a density. However, even though the problem is convex it remains hard to solve because in first or second-order optimization algorithms, computing the gradient or Hessian at a current iterate $\mathbf{u}_d = (u_\alpha)$ requires evaluating integrals of the form

$$\int_\Omega \mathbf{x}^\alpha \exp\left(\sum_{|\alpha| \leq d} u_\alpha \mathbf{x}^\alpha \right) d\lambda(\mathbf{x}), \quad \alpha \in \mathbb{N}_d^n$$

which is a difficult task in general, except perhaps in small dimension $n = 1$ or 2.

3 The mean squared error approach

In this section we assume that the unknown density u is an element of $L^2(\Omega, \lambda)$, and we introduce our mean squared error, or L^2-norm approach, for density approximation.

3.1 Density approximation as an unconstrained problem

We now restrict u_d to be a polynomial, i.e.

$$\mathbf{x} \mapsto \quad u_d(\mathbf{x}) := \sum_{|\alpha| \leq d} u_\alpha \mathbf{x}^\alpha$$

for some vector of coefficients $\mathbf{u} = (u_\alpha) \in \mathbb{R}^{s(d)}$. We first show how to obtain a polynomial estimate $u_d^* \in \mathbb{R}[\mathbf{x}]_d$ of u satisfying (2) by solving an unconstrained optimization problem.

Let $\mathbf{z}$ denote the sequence of moments of λ on Ω, i.e., $\mathbf{z} = (z_\alpha)$, $\alpha \in \mathbb{N}^n$, with

$$z_\alpha = \int_\Omega \mathbf{x}^\alpha d\lambda, \qquad \forall \, \alpha \in \mathbb{N}^n,$$

and let $\mathbf{M}_d(\mathbf{z})$ denote the moment matrix of order d of λ. This matrix is easily computed since the moments of λ are known.

Consider the unconstrained optimization problem

$$\min_{u_d \in \mathbb{R}[\mathbf{x}]_d} \|u - u_d\|_2^2 \left(= \int_\Omega (u - u_d)^2 d\lambda \right). \tag{5}$$

Proposition 1 *Let $\Omega \subset \mathbb{R}^n$ have a nonempty interior and let $\lambda(O) > 0$ for some open set $O \subset \Omega$. A polynomial $u_d^* \in \mathbb{R}[\mathbf{x}]_d$ is an optimal solution of problem (5) if and only if its vector of coefficients $\mathbf{u}_d^* \in \mathbb{R}^{s(d)}$ is an optimal solution of the unconstrained quadratic optimization problem*

$$\min_{\mathbf{u}_d \in \mathbb{R}^{s(d)}} \{ \mathbf{u}_d^T \mathbf{M}_d(\mathbf{z}) \mathbf{u}_d - 2 \mathbf{u}_d^T \mathbf{y} \}. \tag{6}$$

Then $\mathbf{u}_d^* := \mathbf{M}_d^{-1}(\mathbf{z})\mathbf{y}$ is the unique solution of (6), and $u_d^* \in \mathbb{R}[\mathbf{x}]_d$ satisfies:

$$\int_\Omega \mathbf{x}^\alpha u_d^* d\lambda = y_\alpha = \int_\Omega \mathbf{x}^\alpha u \, d\lambda, \qquad \forall \alpha \in \mathbb{N}_d^n. \tag{7}$$

Proof: Observe that for every $u_d \in \mathbb{R}[\mathbf{x}]_d$ with vector of coefficients $\mathbf{u}_d \in \mathbb{R}^{s(d)}$,

$$\int_\Omega (u - u_d)^2 \, d\lambda = \int_\Omega u_d^2 \, d\lambda - 2 \int_\Omega \underbrace{u_d \, u \, d\lambda}_{u_d d\mu} + \int_\Omega u^2 d\lambda$$

$$= \mathbf{u}_d^T \mathbf{M}_d(\mathbf{z})\mathbf{u}_d - 2 \mathbf{u}_d^T \mathbf{y} + \int_\Omega u^2 \, d\lambda.$$

The third term on the right handside being constant, it does not affect the optimization and can be ignored. Thus, the first claim follows.

The second claims follows from the well-known optimality conditions for unconstrained, convex quadratic programs and the fact that $\mathbf{M}_d(\mathbf{z})$ is nonsingular because $\mathbf{M}_d(\mathbf{z}) \succ 0$ for all $d \in \mathbb{N}$. Indeed, if $\mathbf{q}^T \mathbf{M}_d(\mathbf{z})\mathbf{q} = 0$ for some $0 \neq \mathbf{q} \in \mathbb{R}^{s(d)}$ then necessarily the polynomial $q \in \mathbb{R}[\mathbf{x}]_d$ with coefficient vector $\mathbf{q}$ vanishes on the open set O, which implies that $q = 0$, in contradiction with $\mathbf{q} \neq 0$.

Finally, let $\mathbf{e}_\alpha \in \mathbb{R}^{s(d)}$ be the vector of coefficients associated with the monomial $\mathbf{x}^\alpha$, $\alpha \in \mathbb{N}_d^n$. from $\mathbf{M}_d(\mathbf{z})\mathbf{u}_d^* = \mathbf{y}$ we deduce

$$y_\alpha = \mathbf{e}_\alpha^T \mathbf{M}_d(\mathbf{z})\mathbf{u}_d^* = \int_\Omega \mathbf{x}^\alpha u_d^* d\lambda$$

which is the desired result. $\qquad \square$

Thus the polynomial $u_d^* \in \mathbb{R}[\mathbf{x}]_d$ minimizing the L^2-norm distance to u coincides with the polynomial approximation due to [29] defined to be a polynomial which satisfies all conditions (2). Note that this is not the case anymore if one uses an L^p-norm distance with $p > 2$.

Next, we obtain the following convergence result for the sequence of minimizers of problem (5), $d \in \mathbb{N}$.

Proposition 2 *Let Ω be compact with nonempty interior and let λ be finite with $\lambda(O) > 0$ for some open set $O \subset \Omega$. Let (u_d^*), $d \in \mathbb{N}$, be the sequence of minimizers of problem (5). Then $\|u - u_d^*\|_2 \to 0$ as $d \to \infty$. In particular there is a subsequence (d_k), $k \in \mathbb{N}$, such that $u_{d_k}^* \to u$, λ-almost everywhere and λ-almost uniformly on Ω, as $k \to \infty$.*

Proof: Since Ω is compact, $\mathbb{R}[\mathbf{x}]$ is dense in $L^2(\Omega, \lambda)$. Hence as $u \in L^2(\Omega, \lambda)$ there exists a sequence $(v_k) \subset \mathbb{R}[\mathbf{x}]$, $k \in \mathbb{N}$, with $\|u - v_k\|_2 \to 0$ as $k \to \infty$. Observe that if $d_k = \deg v_k$ then as $u_{d_k}^*$ solves problem (5), it holds that $\|u - v_k\|_2 \geq \|u - u_{d_k}^*\|_2$ for all k, which combined with $\|u - v_k\|_2 \to 0$ yields the desired result. The last statement follows from [1, Theorem 2.5.2 and 2.5.3]. $\qquad\square$

Note that computing the L^2 norm minimizer u_d^* is equivalent to solving a system of linear equations, whereas computing the maximum entropy estimate requires solving the potentially hard convex optimization problem (3). Moreover, the L^2-convergence $\|u_d^* - u\|_2 \to 0$ in Proposition 2 (and so almost-everywhere and almost-uniform convergence for a subsequence) is much stronger than the weak convergence (4). On the other hand, unlike the maximum entropy estimate, the L^2-approximation $u_d^* \in \mathbb{R}[\mathbf{x}]_d$ is not guaranteed to be nonnegative on Ω, hence it is not necessarily a density. Methods to overcome this shortcoming are discussed in §3.2.

Remark 1 In general the support $K := \operatorname{supp}\mu$ of μ may be a strict subset of $\Omega = \operatorname{supp}\lambda$. In the case where K is not known or its geometry is complicated, one chooses a set $\Omega \supset K$ with a simple geometry so that moments of λ are computed easily. As demonstrated on numerical examples in Section 4, choosing an enclosing frame Ω as tight as possible is crucial for reducing the pointwise approximation error of u_d^* when the degree d is fixed. In the maximum entropy method of §2.3 no enclosing frame $\Omega \supset K$ is chosen. In the L^2-approach, choosing $\Omega \supset K$ is a degree of freedom that sometimes can be exploited for a fine tuning of the approximation accuracy.

Remark 2 From the beginning we have considered a setting where an exact truncated moment vector $\mathbf{y}$ of the unknown density $u \geq 0$ is given. However, usually one only has an approximate moment vector $\tilde{\mathbf{y}}$ and in fact, it may even happen that $\tilde{\mathbf{y}}$ is not the moment vector of a (nonnegative) measure. In the latter case, the maximum of the convex problem (3) is unbounded, whereas the L^2-norm approach always yields a polynomial estimate $u_d^* \in \mathbb{R}[\mathbf{x}]_d$.

Remark 3 If $\tilde{\mathbf{y}}$ is a slightly perturbed version of $\mathbf{y}$, the resulting numerical error in u_d and side effects caused by ill conditioning may be reduced by considering the regularized problem

$$\min_{u_d} \int_\Omega (u - u_d)^2 d\lambda + \epsilon \|\mathbf{u}_d\|^2 \qquad (8)$$

where $\|\mathbf{u}_d\|^2$ is the Euclidean norm of the coefficient vector $\mathbf{u}_d \in \mathbb{R}^{s(d)}$ of $u_d \in \mathbb{R}[\mathbf{x}]_d$, and $\epsilon > 0$ (fixed) is a regularization parameter approximately of the same order as the noise in $\tilde{\mathbf{y}}$. The coefficient vector of an optimal solution $u_d^*(\epsilon)$ of (8) is then given by

$$\mathbf{u}_d^*(\epsilon) = (\mathbf{M}_d + \epsilon I)^{-1} \tilde{\mathbf{y}}. \qquad (9)$$

The effect of small perturbations in $\mathbf{y}$ on the pointwise approximation accuracy of u_d^* for u is demonstrated on some numerical examples in Section 4. However, a more detailed analysis of the sensitivity of our approach for noise or errors in the given moment information is beyond the scope of this paper.

3.2 Density approximation as a constrained optimization problem

As we have just mentioned, the minimizer u_d^* of problem (5) is not guaranteed to yield a nonnegative approximation even if $u \geq 0$ on Ω. As we next see, the function $\mathbf{x} \mapsto u_{d+}^*(\mathbf{x}) := \max[0, u_d^*(\mathbf{x})]$ also converges to u for the L^2-norm but

- it does not satisfy the moments constraints (i.e., does not match the moments of $d\mu = u d\lambda$ up to order d);

- it is not a polynomial anymore (but a piecewise polynomial).

In the sequel, we address the second point by approximating the density with a polynomial nonnegative on Ω, which for practical purposes, is easier to manipulate than a piecewise polynomial. We do not address explicitly the first point, which is not as crucial in our opinion. Note however that at the price of increasing its degree and adding linear constraints, the resulting polynomial approximation may match an (a priori) fixed number of moments.

Adding a polynomial nonnegativity constraint to problem (5) yields the constrained optimization problem:

$$\min_{u_d \in \mathbb{R}[\mathbf{x}]_d} \{ \|u - u_d\|_2^2 \; : \; u_d \geq 0 \text{ on } \Omega \} \tag{10}$$

which, despite convexity, is untractable in general. We consider two alternative optimization problems to enforce nonnegativity of the approximation; the first one considers necessary conditions of positivity whereas the second one considers sufficient conditions for positivity.

Necessary conditions of positivity

Consider the optimization problem:

$$\min_{u_d \in \mathbb{R}[\mathbf{x}]_d} \{ \|u - u_d\|_2^2 \; : \; \mathbf{M}_d(u_d \, \mathbf{z}) \succeq 0 \}, \tag{11}$$

where $\mathbf{z} = (z_\alpha)$ is the moment sequence of λ and $\mathbf{M}_d(u_d \, \mathbf{z})$ is the localizing matrix associated with u_d and $\mathbf{z}$. Observe that problem (11) is a convex optimization problem because the objective function is convex quadratic and the feasible set is defined by linear matrix inequalities (LMIs).

The rationale behind the semidefiniteness constraint $\mathbf{M}_d(u_d \, \mathbf{z}) \succeq 0$ in (11) follows from a result in [19] which states that if $\operatorname{supp} \lambda = \Omega$ and $\mathbf{M}_d(u_d \, \mathbf{z}) \geq 0$ for all d then $u_d \geq 0$ on Ω.

Lemma 1 *If Ω is compact then $P(\Omega)$ is dense in $L^2(\Omega, \lambda)_+$ with respect to the L^2-norm.*

Proof: As Ω is compact the polynomials are dense in $L^2(\Omega, \lambda)$. Hence there exists a sequence $(u_d) \subset \mathbb{R}[\mathbf{x}]$, $d \in \mathbb{N}$, such that $\|u - u_d\|_2 \to 0$ as $d \to \infty$. But then the sequence

(u_d^+), $d \in \mathbb{N}$, with $u_d^+(\mathbf{x}) := \max[0, u_d(\mathbf{x})]$ for all $\mathbf{x} \in \Omega$, also converges for the L^2-norm. Indeed,

$$
\begin{aligned}
\int_\Omega (u - u_d)^2 d\lambda &= \int_{\Omega \cap \{\mathbf{x}: u_d(\mathbf{x}) < 0\}} (u - u_d)^2 d\lambda + \int_{\Omega \cap \{\mathbf{x}: u_d(\mathbf{x}) \geq 0\}} (u - u_d)^2 d\lambda \\
&\geq \int_{\Omega \cap \{\mathbf{x}: u_d(\mathbf{x}) < 0\}} u^2 d\lambda + \int_{\Omega \cap \{\mathbf{x}: u_d(\mathbf{x}) \geq 0\}} (u - u_d)^2 d\lambda \\
&= \int_\Omega (u - u_d^+)^2 d\lambda = \|u - u_d^+\|_2^2.
\end{aligned}
$$

So let $k \in \mathbb{N}$ and $d_k \in \mathbb{N}$ be such that $\|u - u_{d_k}\|_2 < k^{-1}$ and so $\|u - u_{d_k}^+\|_2 < k^{-1}$. As $u_{d_k}^+$ is continuous and Ω is compact, by the Stone-Weierstrass theorem there exists a sequence $(v_{d_k \ell}) \subset \mathbb{R}[\mathbf{x}]$, $\ell \in \mathbb{N}$, that converges to $u_{d_k}^+$ for the supremum norm. Hence $\sup_{\mathbf{x} \in \Omega} |u_{d_k}^+ - v_{d_k \ell}| < k^{-1}$ for all $\ell \geq \ell_k$ (for some ℓ_k). Therefore, the polynomial $w_{d_k} := v_{d_k \ell_k} + k^{-1}$ is positive on Ω and

$$
\begin{aligned}
\|u - w_{d_k \ell_k}\|_2 &\leq \underbrace{\|u - u_{d_k}^+\|_2}_{< k^{-1}} + \|u_{d_k}^+ - w_{d_k \ell_k}\|_2 \\
&\leq k^{-1} + \underbrace{\|u_{d_k}^+ - v_{d_k \ell_k}\|_2}_{< k^{-1} \lambda(\Omega)^{1/2}} + \underbrace{\|v_{d_k \ell_k} - w_{d_k \ell_k}\|_2}_{= k^{-1} \lambda(\Omega)^{1/2}} \\
&\leq k^{-1} + 2 k^{-1} \lambda(\Omega)^{1/2}.
\end{aligned}
$$

Therefore we have found a sequence $(w_{d_k \ell_k}) \subset P(\Omega)$, $k \in \mathbb{N}$, such that $\|u - w_{d_k \ell_k}\|_2 \to 0$ as $k \to \infty$. $\square$

Proposition 3 *Let Ω be compact with nonempty interior. Then problem (11) has an optimal solution u_d^* for every $d \in \mathbb{N}$, and $\|u - u_d^*\|_2 \to 0$ as $d \to \infty$.*

Proof: Fix d and consider a minimizing sequence $(u_\ell) \subset \mathbb{R}[\mathbf{x}]_d$ with $\|u - u_\ell\|_2^2$ monotonically decreasing and converging to a given value as $\ell \to \infty$. We have $\|u_\ell\|_2 \leq \|u\|_2 + \|u - u_\ell\|_2 \leq \|u\|_2 + \|u - u_0\|_2$ for all $\ell \in \mathbb{N}$. Therefore as $\|\cdot\|_2$ defines a norm on the finite dimensional space $\mathbb{R}[\mathbf{x}]_d$ (Ω has nonempty interior) the whole sequence (u_ℓ) is contained in the ball $\{v : \|v\|_2 \leq \|u\|_2 + \|u - u_0\|_2\}$. As the feasible set is closed, problem (11) has an optimal solution.

Let $(u_d^*) \subset \mathbb{R}[\mathbf{x}]$ be a sequence of optimal solutions of problem (11). By Lemma 1, $P(\Omega)$ is dense in $L^2(\Omega)_+$. Thus there exists a sequence $(v_k) \subset \mathbb{R}[\mathbf{x}]_{\geq 0}$, $k \in \mathbb{N}$, with $\|u - v_k\|_2 \to 0$ as $k \to \infty$. As $v_k \geq 0$ on Ω then necessarily $\mathbf{M}_d(v_k \mathbf{z}) \succeq 0$ for all d and all k. In particular $\mathbf{M}_d(v_{k_d} \mathbf{z}) \succeq 0$ where $k_d := \max\{k : \deg v_k \leq d\}$. Therefore $v_{k_d} \in \mathbb{R}[\mathbf{x}]_d$ is a feasible solution of problem (11) which yields $\|u - v_{k_d}\|_2^2 \geq \|u - u_d^*\|_2^2$ for all d. Combining with $\|u - v_{k_d}\|_2 \to 0$ yields the desired result. $\square$

Via sufficient conditions of positivity

Let $\mathbf{b}_d(\mathbf{x}) := (1, x_1, \ldots, x_n, x_1^2, x_1 x_2, \ldots, x_n^d)^T$ denote the standard monomial basis of $\mathbb{R}[\mathbf{x}]_d$. Let Ω be a basic compact semi-algebraic set defined by $\Omega = \{\mathbf{x} \in \mathbb{R}^n \mid g_j(\mathbf{x}) \geq$

$0, j = 1, \ldots, m\}$ for some polynomials $g_j \in \mathbb{R}[\mathbf{x}]$, $j = 1, \ldots, m$ with $d_j = \lceil (\deg g_j)/2 \rceil$. Then, with $d \geq \max_j d_j$, consider the optimization problem:

$$\min_{u_d \in \mathbb{R}[\mathbf{x}]_d} \quad \|u - u_d\|_2^2$$
$$\text{s.t.} \quad u_d(\mathbf{x}) = \mathbf{b}_d(\mathbf{x})^T \mathbf{A}_0 \mathbf{b}_d(\mathbf{x}) + \sum_{j=1}^m \mathbf{b}_{d-d_j}(\mathbf{x})^T \mathbf{A}_j \mathbf{b}_{d-d_j}(\mathbf{x}) g_j(\mathbf{x}), \qquad (12)$$
$$\mathbf{A}_0 \in \mathbb{S}_+^{s(d)}, \ \mathbf{A}_j \in \mathbb{S}_+^{s(d-d_j)}, \quad j = 1, \ldots, m.$$

Since the equality constraints are linear in u_d and the entries of $\mathbf{A}_j$, $j = 0, \ldots, m$, the feasible set of (12) is a convex LMI set. Moreover, the objective function is convex quadratic.

Note that whereas the semidefinite constraint $\mathbf{M}_d(u_d \, \mathbf{z}) \succeq 0$ in (11) was a relaxation of the nonnegativity constraint $u_d \geq 0$ on Ω, in (12) a feasible solution u_d is necessarily nonnegative on Ω because the LMI constraint on u_d is a Putinar certificate of positivity on Ω. However, in problem (12) we need to introduce $m + 1$ auxiliary matrix variables $\mathbf{A}_j$, whereas (11) is an optimization problem in the original coefficient vector $\mathbf{u}_d$ and does not require such a lifting. Thus, problem (11) is computationally easier to handle than problem (12), although both are convex SDPs, which are substantially harder to solve than problem (6).

Proposition 4 *Let Ω be compact with nonempty interior. For every $d \geq \max_j d_j$, problem (12) has an optimal solution u_d^* and $\|u - u_d^*\|_2^2 \to 0$ as $d \to \infty$.*

Proof: With $d \geq \max_j d_j$ fixed, consider a minimizing sequence $(u_d^*) \subset \mathbb{R}[\mathbf{x}]_d$ for problem (12). As in the proof of Proposition 3 one has $\|u_d^*\|_1 \leq \|u_d^*\|_2 \leq \|u\|_2 + \|u - u_0^*\|_2$. As u_d^* is feasible,

$$u_d^*(\mathbf{x}) = \mathbf{b}_d(\mathbf{x})^T \mathbf{A}_0^d \mathbf{b}_d(\mathbf{x}) + \sum_{j=1}^m \mathbf{b}_{d-d_j}(\mathbf{x})^T \mathbf{A}_j^d \mathbf{b}_{d-d_j}(\mathbf{x}) g_i(\mathbf{x}),$$

for some real symmetric matrices $\mathbf{A}_j^d \succeq 0$, $j = 0, \ldots, m$. Rewriting this as

$$u_d^*(\mathbf{x}) = \langle \mathbf{A}_0^d, \mathbf{b}_d(\mathbf{x}) \mathbf{b}_d(\mathbf{x})^T \rangle + \sum_{j=1}^m \langle \mathbf{A}_j^d, \mathbf{b}_{d-d_j}(\mathbf{x}) \mathbf{b}_{d-d_j}(\mathbf{x})^T \rangle,$$

and integrating with respect to λ yields

$$\langle \mathbf{A}_0^d, \mathbf{M}_d(\mathbf{z}) \rangle + \sum_{j=1}^m \langle \mathbf{A}_j^d, \mathbf{M}_{d-d_j}(g_j \, \mathbf{z}) \rangle = \int_\Omega u_d^* \, d\lambda \leq \|u_d^*\|_1 \leq a,$$

with $a := \|u_d^*\|_1 \leq \|u\|_2 + \|u - u_0^*\|_2$. Hence, for every d,

$$\langle \mathbf{A}_0, \mathbf{M}_d(\mathbf{z}) \rangle \leq a \quad \text{and} \quad \langle \mathbf{A}_j^d, \mathbf{M}_{d-d_j}(g_j \, \mathbf{z}) \rangle \leq a, \quad j = 1, \ldots, m.$$

As $\mathbf{M}_d(\mathbf{z}) \succ 0$, $\mathbf{M}_{d-d_j}(g_j \, \mathbf{z}) \succ 0$, and $\mathbf{A}_j^d \succeq 0$, $j = 1, \ldots, m$, we conclude that all matrices $\mathbf{A}_j^d$ are bounded. Therefore the minimizing sequence $(u_d^*, (\mathbf{A}_j^d))$ belongs to a closed bounded set and as the mapping $v \mapsto \|u - v\|_2^2$ is continuous, an optimal solution exists.

From Lemma 1 there exists $(v_k) \subset \mathbb{R}[\mathbf{x}]_{\geq 0}$, $k \in \mathbb{N}$, such that $\|u - v_k\|_2^2 \to 0$ as $k \to \infty$. Using properties of norms, $\|u - v_k\|_2 - k^{-1} \leq \|u - (v_k + k^{-1})\|_2 \leq \|u - v_k\|_2 + k^{-1}$, and so $\|u - (v_k + k^{-1})\|_2 \to 0$ as $k \to \infty$. Moreover, as $v_k + k^{-1}$ is strictly positive on Ω, by Putinar's Positivstellensatz [26], there exists d_k such that

$$v_k(\mathbf{x}) = \mathbf{b}_{d_k}(\mathbf{x})^T \mathbf{A}_0 \mathbf{b}_{d_k}(\mathbf{x}) + \sum_{j=1}^{m} \mathbf{b}_{d_k - d_j}(\mathbf{x})^T \mathbf{A}_j \mathbf{b}_{d_k - d_j}(\mathbf{x}) g_j(\mathbf{x}), \quad \forall \mathbf{x},$$

for some real matrices $\mathbf{A}_j \succeq 0$, $j = 0, \ldots, m$. So letting $d_k^+ = \max[\deg v_k, d_k]$, the polynomial $v_k + k^{-1}$ is a feasible solution of problem (12) whenever $d \geq d_k^+$ and with value $\|u - (v_k + k^{-1})\|_2^2 \geq \theta_{d_k^+}$. Hence $\|u - v_{d_k^+}\|_2^2 \to 0$ as $k \to \infty$, and by monotonicity of the sequence $(\|u - u_d^*\|_2^2)$, $d \in \mathbb{N}$, the result follows. $\qquad\square$

Remark 4 Since Ω is compact, Proposition 3 and 4 imply that minimizers of the two constrained L^2 norm minimization problems (11) and (12) converge almost uniformly to the unknown density u as in the unconstrained case.

Remark 5 Our approach can handle quite general sets Ω and K as support and frame for the unknown measure in the unconstrained and constrained cases, the only restriction being that (i) Ω and K are basic compact semi-algebraic sets, and (ii) one can compute all moments of λ on Ω. In contrast, to solve problem (3) by local minimization algorithms using gradient and possibly Hessian information, integrals of the type

$$\int_\Omega \mathbf{x}^\alpha \exp \left(\sum_\beta u_\beta \mathbf{x}^\beta \right) d\lambda(\mathbf{x})$$

must be evaluated. Such evaluations may be difficult as soon as $n \geq 3$. In particular in higher dimensions, cubature formulas for approximating such integrals are difficult to obtain if Ω is not a box or a simplex.

4 Numerical experiments

In this section, we demonstrate the potential of our method on a range of examples. We measure the approximation error between a density u and its estimate u_d by the average error

$$\bar{\epsilon}_d := \int_\Omega |u(x) - u_d(x)| d\lambda(x)$$

and the maximum pointwise error

$$\hat{\epsilon}_d := \max_{x \in \Omega} |u(x) - u_d(x)|.$$

In some examples we also consider $\bar{\epsilon}_d^o$ and $\hat{\epsilon}_d^o$, the respective errors on particular segments of the interior of Ω. We compare the performance of our approach to the maximum entropy estimation from §2.3. Both methods are encoded in Matlab. We implemented

the mean squared error (L^2) minimization approach for the standard monomial basis, which results in solving a linear system in the unconstrained case. In the constrained case we apply SeDuMi to solve the resulting SDP problem. In both cases, the numerical stability can be improved by using an orthogonal basis (such as Legendre or Chebychev polynomials) for $\mathbb{R}[x]$. In order to solve the unconstrained, concave optimization problem (3), we apply the Matlab Optimization Toolbox command `fminunc` as a black-box solver. The observed performance of the maximum entropy estimation (MEE) method may be improved when applying more specialized software.
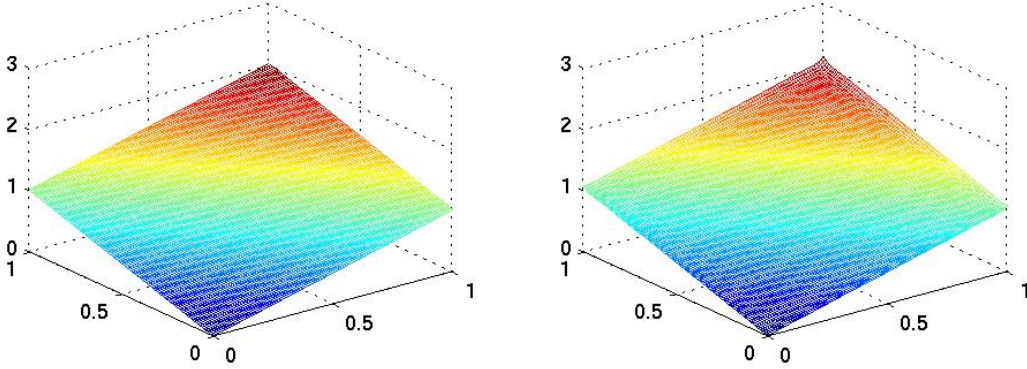

Figure 1: Degree 10 approximation of the density from exact moment vector (left) vs. perturbed moment vector (right).

Example 1 *First we consider the problem of retrieving the density $u(\mathbf{x}) = x_1 + x_2$ on $[0,1]^2$ given its moments $y_\alpha = \int_\Omega \mathbf{x}^\alpha u(x)d\mathbf{x} = \frac{1}{(\alpha_1+1)(\alpha_2+2)} + \frac{1}{(\alpha_1+2)(\alpha_2+1)}$, which was considered as a test case in [24]. This example is a priori very favorable for our technique since the desired u is a polynomial itself. When solving problem (5) for $d \in \{3,5,10\}$ we do obtain the correct solution $\mathbf{u}_d^* = (0,1,1,0,\ldots)^T$ in less than 0.1 secs. Thus, the pointwise approximation is much better than the one achieved in [24]. Moreover, $u_d^* \geq 0$ without adding LMI constraints.*

The question arises of how the polynomial approximation behaves if the moment vector contains some noise, i.e. if it does not exactly coincide with the moment vector of the desired u. We solve problem (5) for $\tilde{\mathbf{y}} := \mathbf{y} + \epsilon$ where the maximal relative componentwise perturbation ϵ between $\mathbf{y}$ and $\tilde{\mathbf{y}}$ is less than 3%. The pointwise error between $u_{10}^*(\mathbf{y})$ and the solution for the perturbed moment vector $u_{10}^*(\tilde{\mathbf{y}})$ is sufficiently small, as pictured in Figure 1.

Example 2 *Next we consider recovering the function $u : [-1,1] \to \mathbb{R}$ with $u(x) = |x|$ as a first example of a nondifferentiable function. In a first step we solve problem (5) for the exact moment vector with entries $y_k := \int_{-1}^1 |x|x^k dx = \frac{1+(-1)^k}{k+2}$, $k = 0,\ldots,d$ corresponding to the density u. The resulting estimates u_d^* for $d \in \{20,30,50\}$ provide a highly accurate pointwise approximation of u on the entire domain, as reported in Table 1 and pictured in Figure 2 (left).*

14

d	problem	$\bar{\epsilon}_d$	$\hat{\epsilon}_d$
20	(5)	0.0031	0.0296
30	(5)	0.0022	0.0265
50	(5)	0.0021	0.0251

Table 1: Estimating the density $u(x) = |x|$ from an exact moment vector.

d	problem	$\bar{\epsilon}_d$	$\hat{\epsilon}_d$	$\bar{\epsilon}_d^o$	$\hat{\epsilon}_d^o$
10	(5)	0.0252	0.2810	0.0207	0.1358
20	(5)	0.0244	0.7934	0.0142	0.0952
30	(5)	0.0237	1.0956	0.0112	0.1024
30	(12)	0.0176	0.4705	0.0106	0.0937
50	(5)	0.0236	1.4591	0.0088	0.1023
50	(12)	0.0206	0.6632	0.0118	0.0979

Table 2: Estimating the density $u(x) = |x|$ from a perturbed moment vector.

In a second step we consider a perturbed moment vector $\tilde{\mathbf{y}}$ as input. When solving problem (5), we observe that both errors $\hat{\epsilon}_d$, $\bar{\epsilon}_d$ on the interior of $[-1, 1]$ – in particular at the nondifferentiable point $x = 0$ – decrease for increasing d, whereas the pointwise error increases at the boundary of the domain. Although providing good approximations for u on the entire interior of $[-1, 1]$, the estimates u_{30}^* and u_{50}^* take negative values at the boundary $\{-1, 1\}$. This is circumvented by solving problem (12). The new estimates are globally nonnegative while their approximation accuracy is only slightly worse than in the unconstrained case, as reported in Table 2 and pictured in Figure 2 (right).

Example 3 *Consider the functions $u_1, u_2 : [-1, 1] \to \mathbb{R}$, $u_1(x) = |x|^{\frac{1}{2}}$, $u_2(x) = |\frac{1}{4} - x^2|^{\frac{1}{2}}$, which are both not locally Lipschitz. Applying mean squared error minimization for $d \in \{20, 50\}$ to the exact moment vectors yields accurate polynomial approximations for both functions on their entire domains, even at the boundary and at the points where the functions are not locally Lipschitz, cf. Figure 3.*

4.1   Recovering geometric objects

One of the main applications of density estimation from moments is the shape reconstruction of geometric objects in image analysis. There has been extensive research on this topic, c.f. [29, 20] and the references therein. The reconstruction of geometric objects is a particular case of density estimation when $u = I_K$, i.e. the desired density u is the indicator function of the geometric object $K \subset \Omega \subset \mathbb{R}^n$. Its given moments $y_\alpha = \int_\Omega \mathbf{x}^\alpha I_K(\mathbf{x}) d\lambda(\mathbf{x}) = \int_K \mathbf{x}^\alpha d\lambda(\mathbf{x})$ do not depend on the frame Ω. However, Ω does enter when computing the matrix $\mathbf{M}_d(\mathbf{z})$ in problem (6). As indicated in Remark 1 and demonstrated below, the choice of the enclosing frame Ω for K is crucial for the pointwise
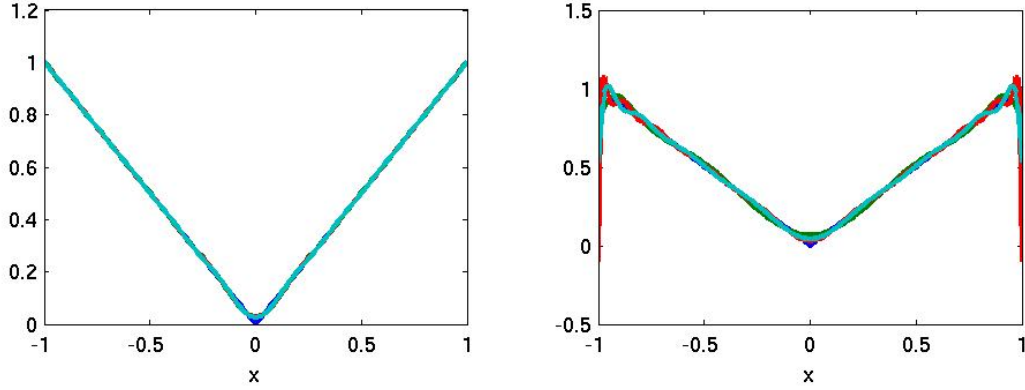
Figure 2: Mean squared error minimizers for $u(x) = |x|$ from exact (left) and perturbed (right) moments. Blue: u, green: u_{20}, red: u_{30}, cyan: u_{50}.
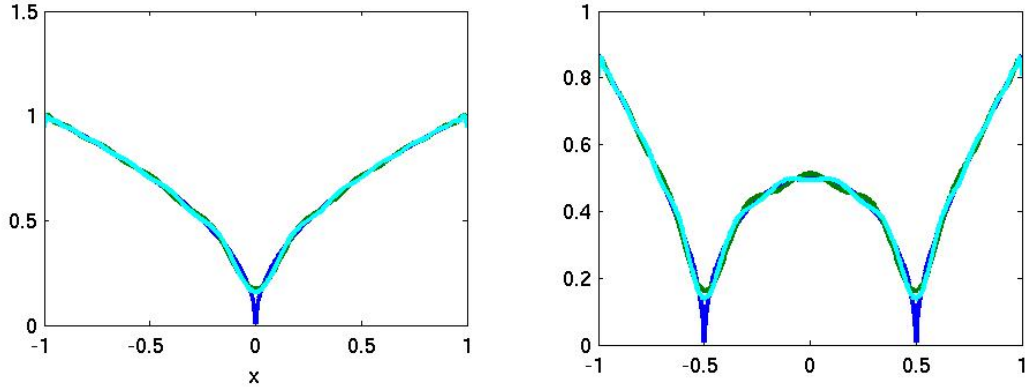

Figure 3: Mean squared error minimizers for u_1 (left) and u_2 (right) from exact moments. Blue: u, green: u_{20}, cyan: u_{50}.

approximation accuracy for a fixed order d. Since u has a special structure, we derive an estimate K_d for K by choosing a superlevel set of u_d^* as proposed in [29]:

$$K_d := \big\{ \mathbf{x} \in \mathbb{R}^n \mid u_d^*(\mathbf{x}) \geq \frac{1}{2} \big\}.$$

Example 4 *A first object we recover is shaped like the letter E as in [29]. We determine the mean squared error minimizer for the same moment vector $\mathbf{y}$ with $d \in \{3, 5, 8, 10\}$, but for three different frames Ω. As pictured in Figure 4, we are able to reconstruct the estimates K_d derived in [29] when Ω is chosen tight. Moreover, we observe that this good approximation of the severely nonconvex set K and its discontinuous indicator function for a small number of moments d depends heavily on the choice for Ω. The wider the frame, the worse gets the approximation accuracy of the truncated estimate. Applying maximum entropy estimation for $d \in \{3, 5, 8\}$ yields density estimates of comparable accuracy than mean squared error minimization, cf. Figure 5. However, the computational effort is*
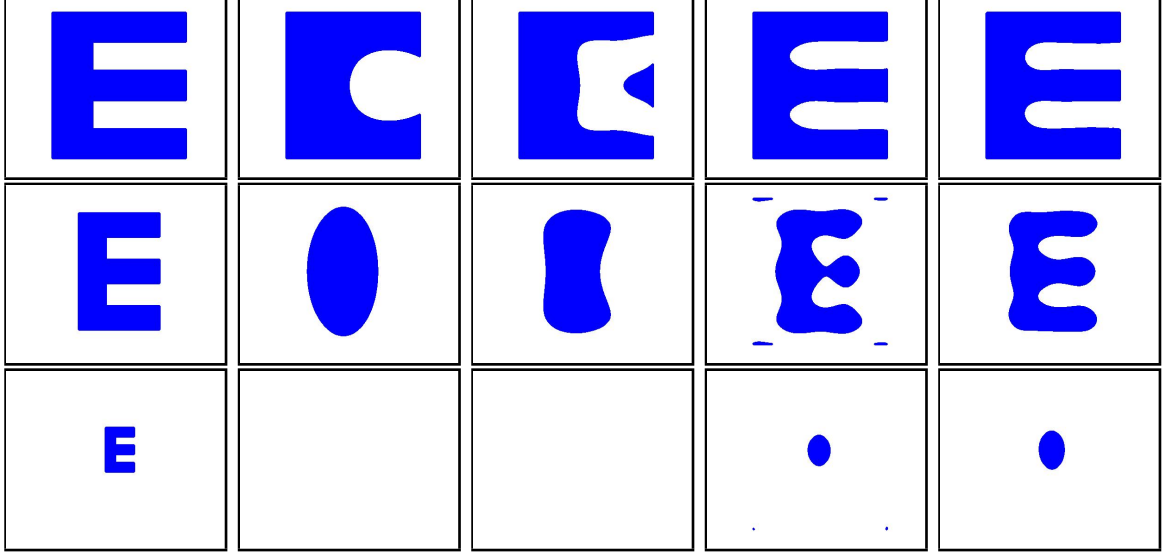
16

Figure 4: Recovering the letter E with mean squared error minimization; tight frames (top) vs. loose frames (bottom); original (left) vs. estimates for degrees $d \in \{3, 5, 8, 10\}$ (from left to right).
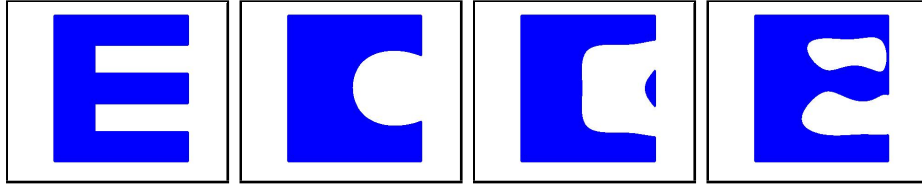

Figure 5: Recovering the letter E with maximum entropy estimation; original (left) vs. estimates for degrees $d \in \{3, 5, 8\}$ (from left to right).

much larger: for $d \in \{3, 5, 8\}$, the computational times for maximum entropy estimation are 82, 902 and 4108 seconds, respectively, whereas the mean squared error minimizer can be determined in less than one second for these values of d.

Example 5 *Secondly we consider recovering an F-shaped set, a less symmetric example than the E-shaped set of Example 4. As previously, we observe that the approximation accuracy of K_d relies heavily on the frame Ω. Even though the set K has a complicated geometry, K_{10} approximates K accurately if Ω is chosen sufficiently tight, cf. Figure 6.*

Example 6 *Consider approximating $K := \{\mathbf{x} \in \mathbb{R}^2 \mid x_1(x_1^2 - 3x_2^2) + (x_1^2 + x_2^2)^2 \geq 0\}$, a nonconvex region enclosed by a trefoil curve. Since this curve is of genus zero, its moment vector $\mathbf{y}$ can be determined exactly. Again, we need to choose an appropriate frame $\Omega \supset K$. The results for $\Omega = B(0, 1)$ and $d \in \{3, 5, 8, 10\}$ are pictured in Figure 7.*
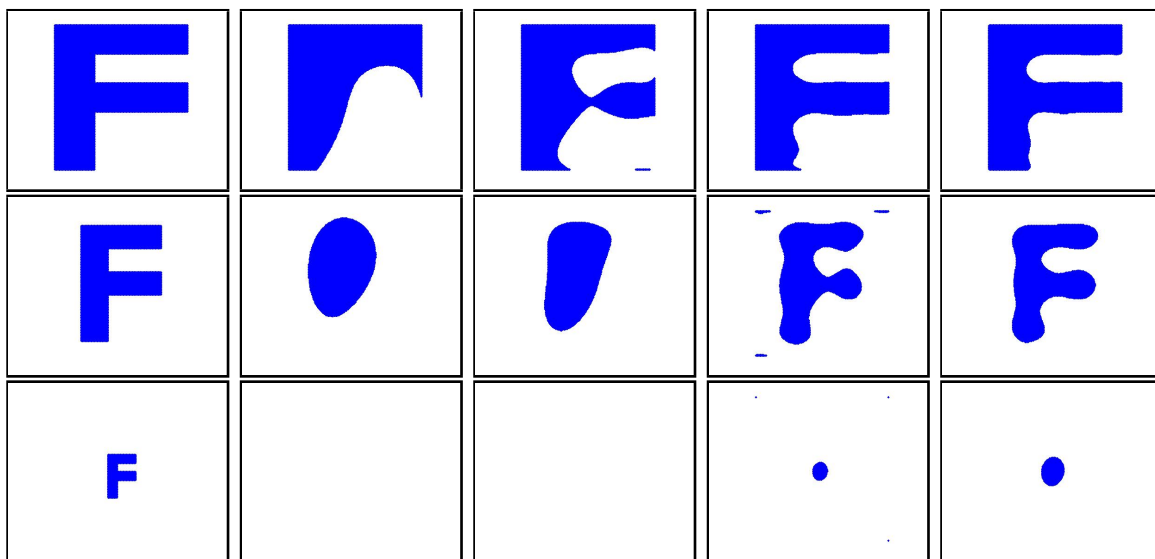
Figure 6: Recovering the letter F with mean squared error minimization; tight frames (top) vs. loose frames (bottom); original (left) vs. estimates for degrees $d \in \{3, 5, 8, 10\}$ (from left to right).
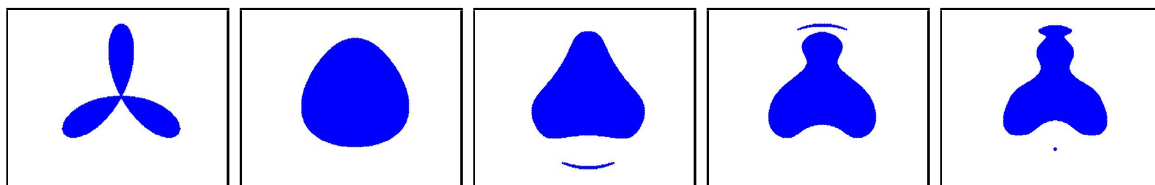

Figure 7: Trefoil region K (left) and $((K_d \setminus K) \cup (K \setminus K_d))$ for $d \in \{3, 5, 8, 10\}$ (from left to right).

4.2   Approximating solutions of differential equations

Another important class of moment problems arises from the numerical analysis of ordinary and partial differential equations. Solutions of certain nonlinear differential equations can be understood as densities of measures associated with the particular differential equation. We obtain an approximate solution from the moment vector of this measure by solving an SDP problem. Approaches for deriving moment vectors associated with solutions of nonlinear differential equations have been introduced in [22, 10] and are omitted here. We assume that the moment vector of a measure whose density is a solution of the respective differential equation is given in the following examples.

Example 7 *Given the moment vectors of a solution (u, v) of the following reaction-*
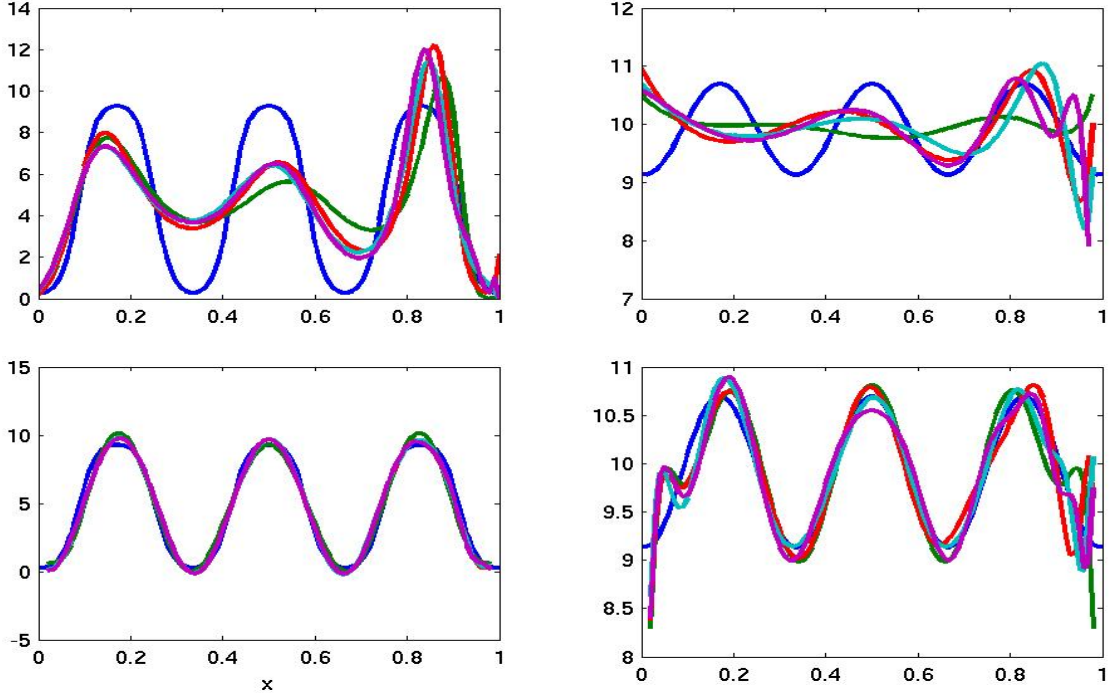
Figure 8: Estimates for u (left) and v (right) of the reaction-diffusion equation via maximum entropy estimation (above) and mean squared error minization (below). Blue: u, green: u_{10}^*, red: u_{20}^*, cyan: u_{30}^*, magenta: u_{50}^*.

diffusion equation [23]:

$$
\begin{array}{ll}
\frac{1}{20}\,u'' + \frac{1}{9}\left(35 + 16u - u^2\right)\,u - u\,v = 0, & \\
4v'' - \left(1 + \frac{2}{5}v\right)\,v + u\,v = 0, & \\
u'(0) = u'(5) = v'(0) = v'(5) = 0, & \\
0 \le u, v \le 14 & on\ [0,5],
\end{array}
$$

we apply both maximum entropy estimation and mean squared error minimization to approximate the desired solution. For the numerical results, see Table 3 and Figure 8. We observe that the mean squared error minimizers provide accurate pointwise approximations for (u,v) on the entire domain, whereas the maximum entropy estimates provides a fairly accurate pointwise approximation on some segment of the domain only. Moreover, the mean squared error minimizers are obtained extremely fast as solutions of linear systems compared with the maximum entropy estimates. Thus, in this example, mean squared error minimization is clearly superior to maximum entropy when estimating densities from moments.

Example 8 *Given the moment vector of the nontrivial, positive solution of the Allen-Cahn bifurcation PDE:*

$$
\begin{array}{ll}
u_{xx} + u_{yy} + 22u(1 - u^2) = 0 & on\ [0,1]^2, \\
u = 0 & on\ \partial[0,1]^2, \\
0 \le u \le 1 & on\ [0,1]^2,
\end{array} \tag{13}
$$

19

d	function	method	$\bar{\epsilon}_d(M)$	$\hat{\epsilon}_d$	$\bar{\epsilon}_d^o$	$\hat{\epsilon}_d^o$	time (sec.)
10	u	MEE	2.0536	4.1063	2.1679	4.1063	38
30	u	MEE	1.7066	3.4775	1.7974	3.4775	2
50	u	MEE	1.6978	3.3858	1.7792	3.3858	820
10	u	L^2	0.4611	1.4506	0.4619	1.4506	0.1
30	u	L^2	0.3942	1.2431	0.4101	1.2431	0.1
50	u	L^2	0.3941	1.2173	0.4004	1.2173	0.1
10	v	MEE	0.5782	2.0982	0.5205	1.0806	221
30	v	MEE	0.5978	9.6060	0.4592	1.2356	645
50	v	MEE	0.6853	23.5993	0.4117	1.2973	3306
10	v	L^2	0.3429	5.4767	0.1765	0.5617	0.1
30	v	L^2	0.3286	12.4501	0.1024	0.6253	0.1
50	v	L^2	0.3744	14.2223	0.1454	0.5907	0.1

Table 3: Approximation accuracy of maximum entropy estimation (MEE) and mean squared error (L^2) optima for the reaction-diffusion equation.

d	method	$\bar{\epsilon}_d$	$\bar{\epsilon}_d^o$	time (sec.)
4	MEE	1.3e-2	1.1e-2	566
6	MEE	9.7e-3	8.9e-3	2489
4	L^2	1.8e-3	1.4e-3	0.1
6	L^2	4.6e-4	2.8e-4	0.1
10	L^2	5.5e-4	2.1e-4	0.1
12	L^2	5.7e-4	2.1e-4	0.1

Table 4: Approximation accuracy of maximum entropy estimation (MEE) and mean squared error (L^2) optima for the Allen-Cahn bifurcation PDE.
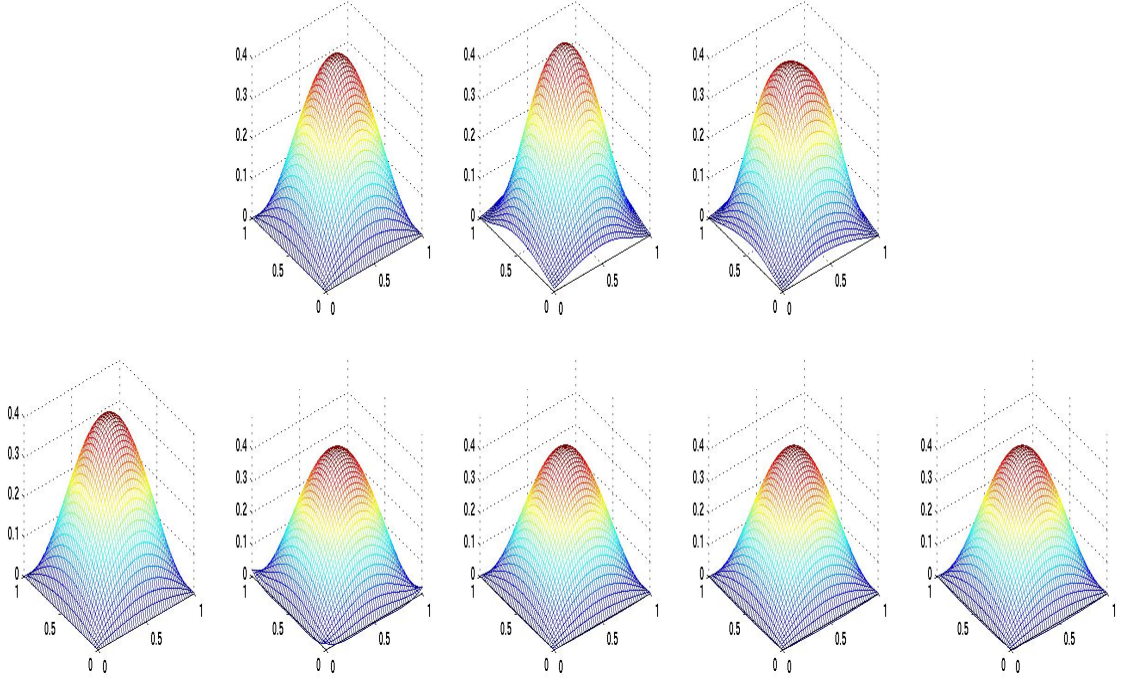
Figure 9: Estimates for solution of the Allen-Cahn bifuraction PDE by maximum entropy estimation (above) and mean squared error minization (below). Exact solution (left) vs estimates for $d \in \{4, 6\}$ and $d \in \{4, 6, 10, 12\}$, respectively.

we apply both maximum entropy and mean squared error minization. The numerical results are reported in Table 4. The approximation accuracy for both methods is comparable, with mean squared error minimization being slightly more precise. However, applying maximum entropy estimation is limited for this problem as the cases $d > 6$ are numerically too heavy to be solved in reasonable time. Mean squared error minimization yields increasingly better estimates for the desired solution within seconds, as pictured in Figure 9.

Example 9 *Given the moment vector of the viscosity solution of the classical eikonal*

d	method	$\bar{\epsilon}_d$	$\bar{\epsilon}_d^o$	time (sec.)
3	MEE	5.4e-2	5.1e-2	20
6	MEE	1.9e-2	1.9e-2	2192
3	L^2	2.8e-2	2.6e-2	0.1
6	L^2	2.8e-2	2.6e-2	0.1
10	L^2	1.4e-2	1.4e-2	0.2

Table 5: Approximation accuracy of maximum entropy estimation (MEE) and mean squared error (L^2) minization for the eikonal PDE.
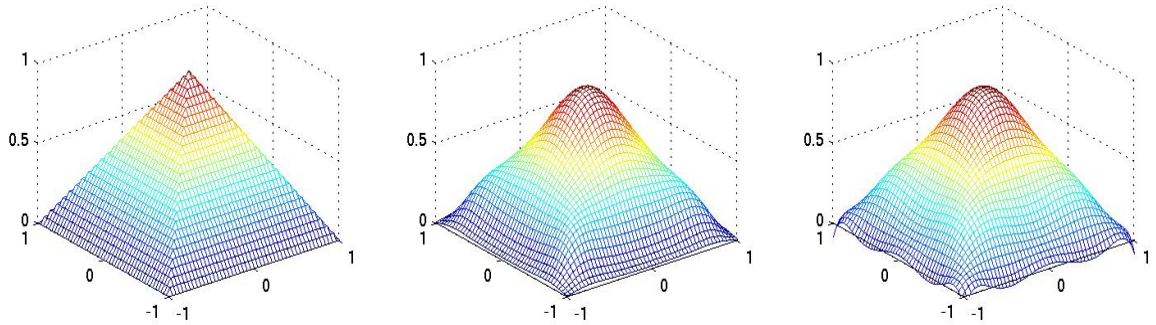
Figure 10: Estimates for solution (left) of eikonal PDE by maximum entropy estimation (center) for $d = 6$ and mean squared error minimization (right) for $d = 10$.

d	problem	$\bar{\epsilon}_d$	$\hat{\epsilon}_d$	time (sec.)
10	(5)	0.08	0.50	0.1
10	(12)	0.11	0.56	0.9
50	(5)	0.05	0.50	0.1
50	(12)	0.08	0.54	2.7
100	(5)	0.05	0.50	0.3
100	(12)	0.07	0.55	37.5

Table 6: Mean squared error minimization for $u(x) = I_{[0.5,1]}(x)$.

PDE:

$$\begin{aligned} u_x^2 + u_y^2 - 1 = 0 & \quad on \ [0,1]^2, \\ u = 0 & \quad on \ \partial[0,1]^2, \end{aligned} \qquad (14)$$

we apply both maximum entropy and mean squared error minization. Both methods provide better pointwise approximates for increasing degrees d, cf. Table 5 and Figure 10. Again, the advantage of using mean squared error minimization is its negligible computation time.

4.3 Approximating indicator functions

In all examples discussed so far, the polynomial approximates u_d^* of solution u obtained by solving the unconstrained problem (5) have been nonnegative on Ω. It was therefore not necessary to solve the constrained problems (11) or (12).

Example 10 *As a first simple example where solving one of the constrained problems is required, consider $u : [0,1] \to \mathbb{R}, u(x) = I_{[0.5,1]}(x)$ and its given vector of moments. We solve problems (5) and (12) for $d \in \{10, 50, 100\}$. As illustrated in Figure 11, solving problem (12) provides globally nonnegative estimates for u. This comes at the price of decreasing approximation accuracy compared to solving the unconstrained problem, since the feasible set of problem (12) is a subset of the feasible set of (5), cf. Table 6.*
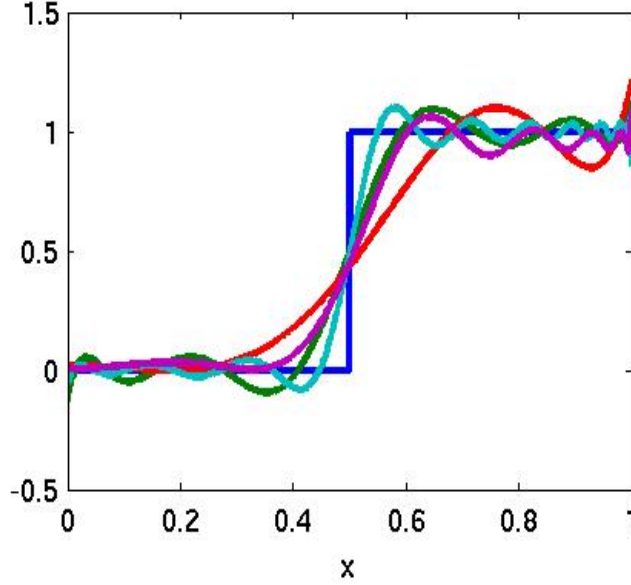
22

Figure 11: Unconstrained and constrained mean squared error minimization estimates. Blue: u, green: u_{10}^* unconstrained, red: u_{10}^* constrained, cyan: u_{100}^* unconstrained, magenta: u_{100}^* constrained.

5 Orthogonal bases and the Gibbs effect

Density estimation via L^2 norm minimization over the truncated space of polynomials is a special case of approximating a function belonging to a complicated functional space by a linear combination of basis elements of an easier, well-understood functional space. This general setting has been considered for both real and complex functional spaces. In the following we discuss the particular case when the basis of the easier functional space is orthogonal, and also effects resulting from truncating the infinite dimensional series expansion of the unknown function.

5.1 The complex case

The classic, complex analogue of the real, inverse moment problem discussed in the previous sections is the problem of approximating a periodic function u by a trigonometric polynomial. An orthonormal basis for the space of trigonometric polynomials is given by $(e^{-ikx})_{k \in \mathbb{Z}}$. It provides the Fourier series expansion:

$$u(x) = \sum_{k \in \mathbb{Z}} y_k e^{ikx},$$

where $y_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} u(x) e^{-ikx} dx$ in the univariate case, or

$$u(x, y) = \sum_{k,l \in \mathbb{Z}} y_{k,l} e^{ikx} e^{ily},$$

23

where $y_{k,l} = \frac{1}{4\pi^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} u(x,y)e^{-ikx}e^{-ily}dx\,dy$ in the bivariate case. One obtains a trigonometric polynomial approximation for u when truncating this series expansion at some $d \in \mathbb{N}$,

$$u_d(x) = \sum_{k=-d}^{d} y_k e^{ikx}.$$

It is a well-known fact in Fourier theory that

$$\int_0^{2\pi}(u-u_d)^2 dx \to 0 \text{ for } d \to \infty.$$

The Fourier approximation for a periodic function is therefore the trigonometric analogue of the real polynomial approximation for a function obtained by mean squared error minimization in our approach. As in the real case we have almost uniform convergence. However, u_d does not converge to u uniformly if u piecewise continuously differentiable with jump discountinuities, due to an effect known as the Gibbs phenomenon. It states that the truncated approximation shows a near constant overshoot and undershoot near a jump discontinuity. This overshoot does not vanish, it only moves closer to the jump for increasing d.

Since the truncated Fourier series is the trigonometric analogue of unconstrained mean squared norm minimizer of problem (5), the question arises of how the trigonometric estimate behaves for a periodic functions when adding nonnegativity constraints as in problems (11) and (12) which aim at preventing the estimate from over- or undershooting near jump discontinuities.

In order to derive a tractable SDP problem, the nonnegativity constraints for the trigonometric polynomial need to be relaxed or tightened to LMI constraints. The difference with the real case will be the moment and localizing matrices being of Toeplitz type in contrast with the Hankel type matrices in the constraints of problem (11).

5.2 The real case

In our discussions in the previous sections we approximated an unknown density by a linear combination of elements of the monomial basis of the space of real polynomials. This choice of a basis for $\mathbb{R}[x]$ has several theoretical and practical shortcomings. For once, it is not an orthogonal basis with respect to the Lebesgue measure, and moreover the moment matrix $\mathbf{M}_d(\mathbf{z})$ in problem (5), whose inverse needs to be computed to determine u_d^*, is severely ill-conditioned [4, 29]. As already pointed out in [29], when choosing an orthogonal basis such as Legendre or Chebychev polynomials for $\mathbb{R}[x]$, the L^2 norm minimizer u_d^* can be determined according to a closed-form formula. Thus, we do not even have to solve a linear system of equations. This is essentially the real analogue of the closed form formula for the coefficients in the Fourier series approximation of periodic functions.

In §5.1 we discussed the Gibbs phenomenon. For reasons outlined there, we expect to observe this effect in the real case as well, when approximating an L^2 integrable function with jump discontinuities. In fact, the function from Example 10 illustrates that.

As shown in Figure 11, we observe an over- and undershoot on both sides of the discontinuity. The amplitude of this overshoot does not decrease for increasing d, but the overshoot moves closer to the jump. When adding the global nonnegativity constraint for the polynomial estimate, the undershoot at the left side disappears. However, it is compensated for by a weaker, overall pointwise approximation accuracy of the estimate. This observation is a first, partial answer to the question raised in the complex case.

6   Conclusion

We introduced an approach for estimating the density of a measure given a finite number of its moments only, in the multivarite case and with no continuity assumption on the density. As an estimate we choose the polynomial minimizing the L^2 norm distance, or mean squared error to the unknown density. We have shown that this estimate is easy to determine by solving a linear system of equations. Moreover, it converges almost uniformly towards the desired density when the degree increases. By minimizing the mean squared error subject to additional linear matrix inequality constraints, which translates to solving a semidefinite program, we obtain a density estimate guaranteed to be nonnegative on the support of the measure. Also in the constrained case, we have shown almost uniform convergence of the nonnegative mean squared error minimizer towards the unknown density, when the degree increases. Mean squared error minimization is often superior to maximum entropy estimation in terms of approximation accuracy and computation time, as demonstrated for a number of examples. Moreover, it is able to handle general, basic, compact semialgebraic sets as support for the unknown measure, which present a challenge to maximum entropy estimation in higher dimension.

References

[1] R.B. Ash. *Real Analysis and Probability*, Academic Press Inc., Boston, 1972.

[2] G.A. Athanassoulis, P.N. Gavriliadis. The truncated Hausdorff moment problem solved by using kernel density functions, Probabilistic Engineering Mechanics 17:273-291, 2002.

[3] G. A. Baker, P. Graves-Morris. Padé approximants. Addison-Wesley, Reading, MA, 1981.

[4] M. Bertero, C. De Mol, E. R. Pike. Linear inverse problems with discrete data. I: General formulation and singular system analysis. Inverse Problems 1(4):301–330, 1985.

[5] J. M. Borwein, A. S. Lewis. On the convergence of moment problems. Trans. Amer. Math. Soc. 325(1):249–271, 1991.

[6] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, D. Picard. Density estimation by wavelet thresholding. Ann. Statist. 24(2):508–539, 1996.

[7] A. Eloyan, S. K. Ghosh. Smooth density estimation with moment constraints using mixture distributions. J. Nonparamet. Stat. 23(2):513–531, 2011.

[8] P.N. Gavriliadis, G.A. Athanassoulis. The truncated Stieltjes moment problem solved by using kernel density functions, Journal of Computational and Applied Mathematics 236:4193-4213, 2012.

[9] R. K. Goodrich, A. Steinhardt. L_2 spectral estimation, SIAM J. Appl. Math. 46(3):417–426, 1986.

[10] D. Henrion, J. B. Lasserre, M. Mevissen. The generalized problem of moments for nonlinear differential equations. Work in progress, 2012.

[11] A. J. Izenman. Recent developments in nonparametric density estimation. J. Amer. Stat. Assoc. 86(413):205–224, 1991.

[12] E. T. Jaynes. Information theory and statistical mechanics, Phys. Rev. Series II 106(4):620–630, 1957.

[13] E. T. Jaynes. Information theory and statistical mechanics II. Phys. Rev. Series II 108(2):171–190, 1957.

[14] V. John, I. Angelov, A. A. Öncül, D. Thévenin. Techniques for the reconstruction of a distribution from a finite number of its moments. Chemical Engineering Science 62(11):2890–2904, 2007.

[15] L. K. Jones, V. Trutzer, On extending the orthogonality property of minimum norm solutions in Hilbert space to general methods for linear inverse problems, Inverse Problems 6(3):379–388, 1990.

[16] G. Kerkyacharian, D. Picard. Density estimation by kernel and wavelet methods: optimality of Besov spaces, Stat. Probab. Letters 18(4):327–336, 1993.

[17] H. J. Landau. Moments in Mathematics. Proc. Symp. Appl. Math., Vol. 37, Amer. Math. Soc., 1987.

[18] J. B. Lasserre. Moments, positive polynomials and their applications. Imperial College Press, London, UK, 2009.

[19] J. B. Lasserre. A new look at nonnegativity on closed sets and polynomial optimization. SIAM J. Optim. 21:864–885, 2011.

[20] S. X. Liao, M. Pawlak. On image analysis by moments. IEEE Trans. Pattern Analysis and Machine Intelligence 18(3):254–266, 1996.

[21] L. R. Mead, N. Papanicolaou. Maximum entropy in the problem of moments. J. Math. Phys. 25(8):2404–2417, 1984.

[22] M. Mevissen, J. B. Lasserre, D. Henrion. Moment and SDP relaxation techniques for smooth approximations of problems involving nonlinear differential equation. Proc. IFAC World Congress on Automatic Control, Milan, Italy, 2011.

[23] M. Mimura. Asymptotic behaviors of a parabolic system related to a planktonic prey and predator model. SIAM J. Appl. Math. 37(3):499-512, 1979.

[24] R. M. Mnatsakanov. Moment-recovered approximations of multivariate distributions: the Laplace transform inversion. Stat. Prob. Letters 81(1):1–7, 2011.

[25] E. Parzen. On estimation of a probability density function and mode. Ann. Math. Stat. 33:1065-1076, 1962.

[26] M. Putinar. Positive polynomials on compact semi-algebraic sets. Indiana Univ. Math. J. 42:969–984, 1993.

[27] S.B. Provost. Moment-Based Density Approximations, The Mathematica Journal 9(4):727-756, 2005.

[28] G. Talenti. Recovering a function from a finite number of moments. Inverse Problems 3(3):501-517, 1987.

[29] M. R. Teague. Image analysis via the general theory of moments. J. Opt. Soc. Amer. 70(8):920-930, 1980.

[30] M. Vannucci. Nonparametric density estimation using wavelets. Discussion Paper 95-26, ISDS, Duke University, 1998.