Mathematical Programming
https://doi.org/10.1007/s10107-024-02072-5

FULL LENGTH PAPER

Series A

Sum-of-squares relaxations for polynomial min–max problems over simple sets

Francis Bach[1,2]

Received: 19 July 2023 / Accepted: 6 February 2024

Abstract
We consider min–max optimization problems for polynomial functions, where a multivariate polynomial is maximized with respect to a subset of variables, and the resulting maximal value is minimized with respect to the remaining variables. When the variables belong to simple sets (e.g., a hypercube, the Euclidean hypersphere, or a ball), we derive a sum-of-squares formulation based on a primal-dual approach. In the simplest setting, we provide a convergence proof when the degree of the relaxation tends to infinity and observe empirically that it can be finitely convergent in several situations. Moreover, our formulation leads to an interesting link with feasibility certificates for polynomial inequalities based on Putinar's Positivstellensatz.

Keywords Polynomial optimization · Sum-of-squares · Min–max problems · Semidefinite programming

Mathematics Subject Classification 90C22 · 11E25

1 Introduction

In this paper, we consider min–max optimization problems of the form

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} g(x, y), \tag{1}$$

where $\mathcal{X}$ and $\mathcal{Y}$ are compact sets and g is a continuous function. Throughout the paper, like [1], we will assume that g is a multivariate polynomial. Among particular cases, a finite set $\mathcal{Y}$ leads to the minimization of the maximum of multivariate

✉ Francis Bach
francis.bach@inria.fr

1 Inria, Paris, France

2 Ecole Normale Supérieure, PSL Research University, Paris, France

Published online: 15 March 2024

polynomials, which is typically not a polynomial function. Thus min–max problems extend the reach of polynomial optimization and have applications in several areas, such as robust optimization [2]. Note that we do not consider saddle-point problems where polynomial optimization has already been studied [3].

We will consider algorithms based on the sum-of-squares principle [4, 5]. This problem has been looked at by [1], which models the function $x \mapsto \max_{y \in \mathcal{Y}} g(x, y)$ by a polynomial, as an upper-bound that is tightly converging as the degree of the approximant increases, but slowly and in most interesting cases non finitely. This bound is then minimized in a two-stage approach, which can deal with a set $\mathcal{Y}$ which can be defined through polynomial inequalities. In this paper, we will need to assume that both sets $\mathcal{X}$ and $\mathcal{Y}$ are "simple", in a sense to be defined in Sect. 3. This includes the regular hypercube, the Boolean hypercube, the unit Euclidean sphere or ball, and all Cartesian products of such sets.[1] However, we will consider a one-stage primal-dual approach that is often finitely convergent (although we currently do not have any provable sufficient conditions).

Paper outline We review SOS relaxations over simple sets in Sect. 3 and present our SOS formulation for the min–max problem in Sect. 4, together with algorithms based on kernels and a convergence proof, while in Sect. 5, we perform illustrative experiments. We start by presenting in Sect. 2 the duality principles that underlie our formulations, which apply beyond polynomials.

2 Primal-dual formulations

We first consider the classical primal-dual formulation of minimization problems, before extending it to min–max problems. In this section, we consider continuous functions (not necessarily polynomials).

2.1 Minimization problems

Given a continuous function f defined on a compact set $\mathcal{X}$, minimizing f can be cast as the minimization of $\int_{\mathcal{X}} f(x) d\mu(x)$ over $\mu \in \mathcal{P}(\mathcal{X})$, the set of probability distributions on $\mathcal{X}$, that is,

$$\min_{x \in \mathcal{X}} f(x) = \min_{\mu \in \mathcal{P}(\mathcal{X})} \int_{\mathcal{X}} f(x) d\mu(x), \tag{2}$$

where the minimizer is any measure supported on the minimizers of f. Introducing the notation $\mathcal{M}(\mathcal{X}, Q)$ for the set of finite measures with values in the cone Q, we can see probability distributions as the elements of $\mathcal{M}(\mathcal{X}, \mathbb{R}_+)$ such that $\int_{\mathcal{X}} d\mu(x) = 1$. Introducing a Lagrange multiplier $c \in \mathbb{R}$ for this linear constraint, we get by convex duality:

[1] This assumption is mostly made to make the developments as simple as possible, but most of our developments would go through for any basic semi-algebraic sets $\mathcal{X}$ and $\mathcal{Y}$ through the use of adapted positivity certificates.

$$\min_{\mu\in\mathcal{P}(\mathcal{X})} \int_{\mathcal{X}} f(x)d\mu(x) = \max_{c\in\mathbb{R}} \inf_{\mu\in\mathcal{M}(\mathcal{X},\mathbb{R}_+)} \int_{\mathcal{X}} f(x)d\mu(x) + c\Big(\int_{\mathcal{X}} d\mu(x) - 1\Big)$$
$$= \max_{c\in\mathbb{R}} c \ \text{ such that } \ \forall x \in \mathcal{X}, \ f(x) - c \geq 0, \tag{3}$$

which is equivalent to finding the largest minorant of f (and thus provides a direct proof of strong duality). As shown in Sect. 3, these two equivalent formulations lead to equivalent SOS relaxations, by replacing non-negative functions by sums-of-squares in Eq. (3), and representing probability measures by their moments and "pseudo"-moments in Eq. (2). We now extend these equivalent formulations to min–max problems.

2.2 Min–max problems

We now consider primal-dual interpretations for the original problem in Eq. (1), akin to Eq. (2) and Eq. (3) in Sect. 2.1 above, for a continuous function $g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.

For the outer minimization problem in $x \in \mathcal{X}$, we consider the probabilistic formulation from Eq. (2), and we thus have the equivalent formulation:

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} g(x, y) = \min_{\mu\in\mathcal{P}(\mathcal{X})} \int_{\mathcal{X}} \Big(\max_{y\in\mathcal{Y}} g(x, y)\Big)d\mu(x).$$

For the inner maximization problem in $y \in \mathcal{Y}$, which is different for every $x \in \mathcal{X}$, we consider probability measures $\nu(\cdot|x) \in \mathcal{P}(\mathcal{Y})$ (the set of probability measures on $\mathcal{Y}$), and apply the same reformulation, to obtain

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} g(x, y) = \min_{\mu\in\mathcal{P}(\mathcal{X})} \max_{\nu:\mathcal{X}\to\mathcal{P}(\mathcal{Y})} \int_{\mathcal{X}}\int_{\mathcal{Y}} g(x, y)d\nu(y|x)d\mu(x). \tag{4}$$

This is now a convex-concave min–max problem in infinite dimensions (while the original one in Eq. (1) is typically not), with a bilinear objective and two convex domains, for which min and max can be swapped as the set $\mathcal{P}(\mathcal{X})$ is compact for the weak topology on measures on a compact set $\mathcal{X}$ [6, Corollary 3.3]. We can now use convex duality to obtain either a minimization problem or a maximization problem.

We have, from Eq. (4), by adding the Lagrange multiplier $c \in \mathbb{R}$ for the constraint $\int_{\mathcal{X}} d\mu(x) = 1$:

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} g(x, y) = \min_{\mu\in\mathcal{P}(\mathcal{X})} \max_{\nu:\mathcal{X}\to\mathcal{P}(\mathcal{Y}),\ c\in\mathbb{R}}$$
$$\int_{\mathcal{X}}\int_{\mathcal{Y}} g(x, y)d\nu(y|x)d\mu(x) + c\Big(1 - \int_{\mathcal{X}} d\mu(x)\Big)$$
$$= \max_{\nu:\mathcal{X}\to\mathcal{P}(\mathcal{Y}),\ c\in\mathbb{R}} c \ \text{ such that } \ \forall x \in \mathcal{X}, \int_{\mathcal{Y}} g(x, y)d\nu(y|x) \geq c, \tag{5}$$

which is a *maximization* problem.

Alternatively, by convex duality, this equal to, introducing in Eq. (4) a Lagrange multiplier $\lambda \in \mathcal{M}(\mathcal{X}, \mathbb{R})$ (set of finite signed measures) for the constraint that $\forall x \in \mathcal{X}$, $\int_{\mathcal{Y}} dv(y|x) = 1$ [7]:

$$\min_{\mu \in \mathcal{P}(\mathcal{X}),\ \lambda \in \mathcal{M}(\mathcal{X}, \mathbb{R})} \ \max_{v : \mathcal{X} \to \mathcal{M}(\mathcal{Y}, \mathbb{R}_+)} \int_{\mathcal{X}} \Big(\int_{\mathcal{Y}} g(x, y) dv(y|x) \Big) d\mu(x) + \int_{\mathcal{X}} \Big(1 - \int_{\mathcal{Y}} dv(y|x) \Big) d\lambda(x),$$

which is equal to

$$\min_{\mu \in \mathcal{P}(\mathcal{X}),\ \lambda \in \mathcal{M}(\mathcal{X}, \mathbb{R})} \int_{\mathcal{X}} d\lambda(x) \ \text{ such that } \ \forall y \in \mathcal{Y},\ \lambda \geq g(\cdot, y)\mu, \tag{6}$$

which is another convex formulation as a *minimization* problem.

Overall we get three formulations which are all equivalent to the original problem (in Sect. 4, our SOS formulation will also have these three equivalent formulations):

- *Minimization*, corresponding to Eqs. (23) and (27) in Sect. 4:

$$\min_{\mu \in \mathcal{P}(\mathcal{X}),\ \lambda \in \mathcal{M}(\mathcal{X}, \mathbb{R})} \int_{\mathcal{X}} d\lambda(x) \ \text{ such that } \ \forall y \in \mathcal{Y},\ \lambda \geq g(\cdot, y)\mu. \tag{7}$$

- *Maximization*, corresponding to Eqs. (24) and (28) in Sect. 4:

$$\max_{v : \mathcal{X} \to \mathcal{P}(\mathcal{Y}),\ c \in \mathbb{R}} c \ \text{ such that } \ \forall x \in \mathcal{X},\ \int_{\mathcal{Y}} g(x, y) dv(y|x) \geq c. \tag{8}$$

- *Saddle-point*, corresponding to Eqs. (22) and (26) in Sect. 4:

$$\min_{\mu \in \mathcal{P}(\mathcal{X})} \ \max_{v : \mathcal{X} \to \mathcal{P}(\mathcal{Y})} \int_{\mathcal{X}} \int_{\mathcal{Y}} g(x, y) dv(y|x) d\mu(x). \tag{9}$$

Non-convex formulation By writing $d\lambda(x) = a(x) d\mu(x)$ for a certain function $a : \mathcal{X} \to \mathbb{R}$, which is only possible for a dense subset of $\mathcal{M}(\mathcal{X}, \mathbb{R})$, we get an equivalent reformulation

$$\min_{\mu \in \mathcal{P}(\mathcal{X}),\ a : \mathcal{X} \to \mathbb{R}} \int_{\mathcal{X}} a(x) d\mu(x) \ \text{ such that } \ \forall (x, y) \in \mathcal{X} \times \mathcal{Y},\ a(x) \geq g(x, y), \tag{10}$$

which is a non-convex formulation because the objective is non-convex. An alternating minimization algorithm starting from a measure μ with full support leads to the global optimum after one minimization with respect to a (leading to $a(x) = \max_{y \in \mathcal{Y}} g(x, y)$), and then one minimization with respect to μ (leading to the minimizer of this function a). When using SOS formulations for these two operations, we exactly obtain the formulation of [1] (see Eq. (17) in Sect. 4.1).

Optimal solutions With c_* being the optimal value of Eq. (1), the optimal measure $\mu \in \mathcal{P}(\mathcal{X})$ is any measure supported on the minimizers of $x \mapsto \max_{y\in\mathcal{Y}} g(x, y)$. The optimal $\nu : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ is such that for all $x \in \mathcal{X}$, $\int_{\mathcal{Y}} g(x, y)d\nu(y|x) \geq c_*$ with equality at any minimizer $x_* \in \mathcal{X}$. Therefore, at all minimizers $x_* \in \mathcal{X}$, we need $\nu(\cdot|x_*)$ to put mass only at maximizers of $y \mapsto g(x_*, y)$, but this is *not* required at other positions. The optimal λ is equal to an optimal μ times $\max_{y\in\mathcal{Y}} g(x, y)$.

Relationship to zero-sum polynomial games There is an interesting parallel between Eq. (4) and zero-sum games of the form

$$\min_{\mu\in\mathcal{P}(\mathcal{X})} \max_{\nu\in\mathcal{P}(\mathcal{Y})} \int_{\mathcal{X}} \int_{\mathcal{Y}} g(x, y)d\nu(y)d\mu(x),$$

where now the measure ν does not depend on x. This lack of dependence makes it easier to solve as shown by [8, Section 5].

3 SOS relaxations for polynomials over simple sets

In this section, we review existing work on minimizing polynomial functions over simple sets, which we cast as minimizing a quadratic form $f(x) = \varphi(x)^\top F\varphi(x)$ for a feature map $\varphi : \mathcal{X} \to \mathbb{R}^m$, where $\mathcal{X}$ is a compact set. While we use specific notations that will make further developments easier to describe, this section follows the classical SOS formulations (see [9, 10] for a thorough review).

We use the denomination "simple set" to refer to a set $\mathcal{X}$ coming with its feature map $\varphi : \mathcal{X} \to \mathbb{R}^m$ *with unit norm*, that is, $\|\varphi(x)\|^2 = 1$ for all $x \in \mathcal{X}$ (for the Euclidean norm), and, which can be represented (potentially after transformation) as a multivariate polynomial (this thus imposes that $\mathcal{X}$ is a subset of $\mathbb{R}^d$ for a specific d).

We will always assume that the constant mapping and the identity mapping $x \mapsto x$ can be obtained as a linear function of $\varphi(x)$ (this will be useful in recovering maximizers in Sect. 3.4). Moreover, we will only need to access the positive-definite kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined as $k(x, y) = \varphi(x)^\top\varphi(y)$ (and not access to the vector φ). Our unit norm normalization on φ translates to $k(x, x) = 1$ for all $x \in \mathcal{X}$.

We assume that the dimension of the span of all $\varphi(x)$, $x \in \mathcal{X}$ is m, while the dimension of the span $\mathcal{V}_\varphi$ of all $\varphi(x)\varphi(x)^\top \in \mathbb{R}^{m\times m}, x \in \mathcal{X}$, is $m' \in [m, m(m+1)/2]$. Finally, we assume we can generate (typically, randomly) m' points $x_1, \ldots, x_{m'}$, such that $\varphi(x_i)\varphi(x_i)^\top, i = 1, \ldots, m'$, is a basis of $\mathcal{V}_\varphi$.

The optimization problem and our solution will be invariant by invertible linear transformations, and we can choose the feature map so that the kernel is as simple as possible (note, however, that in terms of conditioning of the associated numerical linear algebra, some kernels are better than others).

All of our examples will be (subsets of) Euclidean unit spheres or products of Euclidean spheres.

3.1 Examples

We will consider the following sets, feature maps, and kernel functions. Since our relaxations are based on approximating non-negative polynomials as sums-of-squares, that is, positive semi-definite quadratic forms in φ, we describe these SOS polynomials for some instances.

- *Discrete data*: $\mathcal{X} = \{1, \dots, m\}$ with orthonormal features $\varphi : \mathcal{X} \to \mathbb{R}^m$, defined as $\varphi(x)_i = 1_{x=i}$. The corresponding kernel is $k(x, x') = 1_{x=x'}$, with $m' = m$.
- *Trigonometric polynomials on* $[0, 1]$: $\mathcal{X} \in [0, 1]$ with $\varphi(x)_\omega = \frac{1}{(2r+1)^{1/2}} e^{2i\pi\omega x}$ for $\omega \in \{-r, \dots, +r\}$.[2] The kernel is $k(x, x') = \frac{\sin[(2r+1)\pi(x-x')]}{(2r+1)\sin\pi(x-x')}$. This can be equivalently represented in the unit Euclidean sphere in $\mathbb{R}^2$ with the bijection $\theta \mapsto (\cos 2\pi\theta, \sin 2\pi\theta)$, where the corresponding feature map spans all bivariate polynomials of degree r, with a kernel that can be taken to be equal to $k(y, y') = \frac{1}{2^r}(1 + y^\top y')^r$ for $y, y' \in \mathbb{R}^2$ of unit norm (we could construct one with Chebyshev polynomials to get the exact equivalence with the kernel above). We then have $m = 2r + 1$ and $m' = 4r + 1$.
- *Polynomials on* $[-1, 1]$: this is simply the projection of the case above by considering $y \in \mathbb{R}^2$ such that $y_1^2 + y_2^2 = 1$, and only considering functions of y_1. As shown in [11], a polynomial in y_1 which is equal to an SOS polynomial on y_1, y_2 can be written as the sum $u(y_1) + (1 - y_1^2)v(y_1)$ where u and v are univariate SOS polynomials.
- *Hypersphere:* $\mathcal{X} = \{x \in \mathbb{R}^{d+1},\ \|x\|_2^2 = x^\top x = 1\}$, with all functions that are multivariate polynomials of degree r. This corresponds to $m = \binom{d+r}{r} + \binom{d+r-1}{r-1}$ and $m' = \binom{d+2r}{2r} + \binom{d+2r-1}{2r-1}$. We can choose the kernels $k(x, x') = \frac{1}{r+1}\sum_{i=0}^r \left(x^\top x'\right)^i$ or $k(x, x') = \frac{1}{2^r}(1 + x^\top x')^r$. We could also use generalized Legendre polynomials [12] to get better-conditioned kernel matrices.
- *Euclidean ball:* $\mathcal{X} = \{x \in \mathbb{R}^d,\ x^\top x \leq 1\}$ can be seen as the projection of the hypersphere above to the first d dimensions. When obtaining an SOS polynomial on the hypersphere, this translates for the Euclidean ball to a sum $u(x) + (1 - \|x\|_2^2)v(x)$ where u and v are SOS polynomials.
- *Products of one-dimensional spheres* $\subset \mathbb{R}^2 \Leftrightarrow$ *trigonometric polynomials on* $[0, 1]^d \Leftrightarrow$ *regular polynomials on* $[-1, 1]^d$: this is the tensor product of the univariate cases above; the kernel is then $k(y, y') = \prod_{i=1}^d \frac{1}{2^r}(1 + y_i^\top y_i')^r$ for the polynomial representations, or alternatively $k(x, x') = \prod_{i=1}^d \frac{\sin[(2r+1)\pi(x_i - x_i')]}{(2r+1)\sin\pi(x_i - x_i')}$ for trigonometric polynomials. This then corresponds to multivariate polynomials of *maximal*[3] degree $2r$. As shown in [11], a trigonometric SOS polynomial transferred to regular polynomials on $[-1, 1]^d$ leads to a representation of Schmudgen's type [13].

[2] This feature is complex-valued but equivalent real-valued formulations with cosines and sines could be used. Since we only use kernel formulations, we do not need to pursue them explicitly.

[3] For a monomial $X_1^{\alpha_1} \cdots X_d^{\alpha_d}$, its degree is $\alpha_1 + \cdots + \alpha_d$ and its *maximal* degree is $\max\{\alpha_1, \dots, \alpha_d\}$.

- *Boolean hypercube* $\mathcal{X} = \{-1, 1\}^d$: it can be seen as a sub-case of the hypersphere in dimension $d - 1$ and radius $\sqrt{d}$, where quadratic forms are polynomials of degree $2r$. We then have $m = \sum_{i=0}^{r} \binom{d}{i}$.

3.2 Relaxation

The SOS relaxation[4] is obtained by first representing the minimization of f as the maximization of a minorant c of f, that is, such that $f(x) - c \geq 0$ for all $x \in \mathcal{X}$, that is, Eq. (3) in Sect. 2. We then represent non-negative functions as sums-of-squares, that is, a positive semi-definite quadratic form in $\varphi(x)$, thus solving:

$$\max_{c \in \mathbb{R},\ A \succcurlyeq 0} \ c \quad \text{such that } \forall x \in \mathcal{X},\ f(x) = c + \varphi(x)^\top A \varphi(x). \tag{11}$$

In general, the formulation above is a *strengthening* of the minimization problem as $f - c \geq 0$ is implied by $\forall x \in \mathcal{X},\ f(x) = c + \varphi(x)^\top A \varphi(x)$, where $A \succcurlyeq 0$, but not equivalent, except in special cases described in Sect. 3.1.

Equation (11) can be re-written using $\mathcal{V}_\varphi$ the span of all $\varphi(x)\varphi(x)^\top$, $x \in \mathcal{X}$, and its orthogonal subspace $\mathcal{V}_\varphi^\perp$, as, using the representation of f through $f(x) = \varphi(x)^\top F \varphi(x)$, where $F \in \mathbb{R}^{m \times m}$:

$$\begin{aligned} &\max_{c \in \mathbb{R},\ A \succcurlyeq 0} \ c \quad \text{such that} \quad \forall x \in \mathcal{X},\ \operatorname{tr}\left[\varphi(x)\varphi(x)^\top (F - cI - A)\right] = 0 \\ &\quad = \max_{c \in \mathbb{R},\ A \succcurlyeq 0,\ Y \in \mathcal{V}_\varphi^\perp} \ c \quad \text{such that} \quad F - cI - A + Y = 0, \quad \text{by definition of } \mathcal{V}_\varphi^\perp. \end{aligned}$$

We can then optimize out c and A, by noticing that $c \in \mathbb{R}$ is the largest c such that $F + Y \succcurlyeq cI$, leading to the following spectral formulation

$$\max_{Y \in \mathcal{V}_\varphi^\perp} \ \lambda_{\min}(F + Y). \tag{12}$$

Its dual can be written as, using standard semi-definite programming duality [14]:

$$\begin{aligned} \max_{Y \in \mathcal{V}_\varphi^\perp} \ \lambda_{\min}(F + Y) &= \min_{\Sigma \succcurlyeq 0} \ \max_{Y \in \mathcal{V}_\varphi^\perp} \ \operatorname{tr}[\Sigma(F + Y)] \ \text{ such that } \ \operatorname{tr}(\Sigma) = 1 \\ &= \min_{\Sigma \succcurlyeq 0} \ \operatorname{tr}(\Sigma F) \ \text{ such that } \ \operatorname{tr}(\Sigma) = 1,\ \Sigma \in \mathcal{V}_\varphi, \end{aligned} \tag{13}$$

which corresponds to an outer approximation of the convex hull of all $\varphi(x)\varphi(x)^\top$, $x \in \mathcal{X}$, by the set of positive semi-definite matrices such that $\operatorname{tr}(\Sigma) = 1$ and $\Sigma \in \mathcal{V}_\varphi$, which we denote $\widehat{\mathcal{K}}_\varphi$ and which is an outer approximation of $\mathcal{K}_\varphi$, the closure of the convex hull of all $\varphi(x)\varphi(x)^\top$, $x \in \mathcal{X}$. This dual formulation corresponds to (a) replacing the minimization of f by the minimization with respect to a probability

[4] In this paper, we use the term "relaxation" for all our formulations, but, rigorously, they are "strengthenings" when replacing non-negative functions by sums-of-squares, and proper relaxations in their dual formulations, when relaxing moments to pseudo-moments later in this section.

measure on $\mathcal{X}$ of the expectation of f with respect to that measure, as done in Sect. 2.1 in Eq. (2), and (b) characterizing these measures by their expectations of $\varphi\varphi^\top$. Elements of $\mathcal{K}_\varphi$ are moment matrices while elements of $\widehat{\mathcal{K}}_\varphi$ are often referred to as "pseudo-moment" matrices.

3.3 Kernelization

With an explicit description of $\mathcal{V}$, it may be cumbersome to implement the semi-definite program, particularly for larger input dimensions, leading to dedicated codes for each case. This is simpler with kernels, as described below. It allows accessing the function f using only function values, like proposed by [15], with a direct link with positive definite kernels outlined by [16]. As we now show, this corresponds to representing the space $\mathcal{V}$ by a span of finitely many elements, leading to a representation of the moment matrices Σ as a linear combination of rank-one matrices. Note that since we consider a finite-dimensional feature map φ, the kernel representation is exact with sufficiently many points.

We consider m' "well-positioned" points $x_1, \dots, x_{m'} \in \mathcal{X}$, so that $\mathcal{V}_\varphi$ is the span of all $\varphi(x_i)\varphi(x_i)^\top$, $i = 1, \dots, m'$. Quasi-random sequences [17] are natural candidates, in particular, because we will extract below the first m points and also need them to be well-spread to avoid ill-conditioning of the kernel matrices.

For the primal formulation in Eq. (11), the constraint that $\forall x \in \mathcal{X}$, $f(x) = c + \varphi(x)^\top A \varphi(x)$ is equivalently replaced by an equality only on $x_1, \dots, x_{m'}$. This corresponds to checking that two polynomials are equal by checking that they are equal on sufficiently many points.

The dual formulation in Eq. (13) is then equivalent to:

$$\inf_{\alpha \in \mathbb{R}^{m'}} \sum_{i=1}^{m'} \alpha_i f(x_i) \text{ such that } \sum_{i=1}^{m'} \alpha_i = 1, \ \sum_{i=1}^{m'} \alpha_i \varphi(x_i)\varphi(x_i)^\top \succcurlyeq 0,$$

which is only accessing the function f through m' function evaluations. The crucial point is that the vector $\alpha \in \mathbb{R}^{m'}$ is not constrained to have non-negative values (otherwise, the formulation above would lead to $\min_{i \in \{1,\dots,m'\}} f(x_i)$).

If m is the dimension of φ, then from the kernel matrix $K \in \mathbb{R}^{m \times m}$ associated with the first m points, we build the "empirical feature map" as $\tilde{\varphi}(x) = K^{-1/2}(k(x_i, x))_{i \in \{1,\dots,m\}} \in \mathbb{R}^m$, where $K^{-1/2}$ is any inverse square root of $K \in \mathbb{R}^{m \times m}$. This defines an empirical feature matrix $\Phi = LK^{-1/2} \in \mathbb{R}^{m' \times m}$, where $L \in \mathbb{R}^{m' \times m'}$ is the full kernel matrix of all m' points. We then solve, equivalently,

$$\inf_{\alpha \in \mathbb{R}^{m'}} \sum_{i=1}^{m'} \alpha_i f(x_i) \text{ such that } \sum_{i=1}^{m'} \alpha_i = 1, \ \Phi^\top \operatorname{diag}(\alpha) \Phi \succcurlyeq 0, \tag{14}$$

and obtain a solution $\Sigma = \sum_{i=1}^{m'} \alpha_i \varphi(x_i)\varphi(x_i)^\top$. We will see below how to obtain a candidate maximizer $x_* \in \mathcal{X}$ from Σ without the need to compute φ.

Going infinite-dimensional Solving Eq. (14) will lead to the SOS relaxation if f is indeed a quadratic form in $\varphi(x)$. In all our examples, the feature map is finite-dimensional. Still, we can go infinite-dimensional using positive definite kernels corresponding to infinite dimensional feature spaces, such as $k(x, x') = \exp(x^\top x')$, with an additional regularizer. See [16] for more details and convergence analysis.

3.4 Practical algorithms

Solving the SDP The problem in Eq. (14) is a semi-definite program, which can either be solved using generic toolboxes, with complexity $(m')^{3.5}$ [18]. Adding a log-determinant barrier leads to an approximate algorithm with only matrix inversions of size m and m' [16], but no eigenvalue decompositions.

Obtaining rank-one solutions The obtained solution $\alpha \in \mathbb{R}^{m'}$ of Eq. (14) may not lead to a rank-one matrix $\Sigma = \sum_{i=1}^{m'} \alpha_i \varphi(x_i)\varphi(x_i)^\top$ when the minimization problem has several minimizers or the relaxation is not tight. We can obtain a lower-rank solution (and rank-one when the relaxation is tight) by minimizing a random linear function of α over all α that are minimizers of Eq. (14). Rank-minimization heuristics could also be used [19].

Obtaining candidates for x_* Once the vector α is obtained such that the matrix $\Sigma = \sum_{i=1}^{m'} \alpha_i \varphi(x_i)\varphi(x_i)^\top$ has rank one, we can simply obtain the corresponding $x_* \in \mathcal{X}$ exactly as $x_* = \sum_{i=1}^{m'} \alpha_i x_i$. This is only approximate when Σ does not have rank one. See also [20].

3.5 Tightness guarantees

For a small number of cases, we have $\widehat{\mathcal{K}}_\varphi = \mathcal{K}_\varphi$, that is, the relaxation is tight, e.g., for one-dimensional problems or with linear features (modeling quadratic polynomials). Otherwise, we need "hierarchies".

Hierarchies For most cases, the relaxation is not tight, that is, $\widehat{\mathcal{K}}_\varphi \supsetneq \mathcal{K}_\varphi$, but we can see a $2r$-dimensional polynomial as an instance of a polynomial of degree less than $2s$, for $s > r$, and run the algorithm with the kernel corresponding to this larger dimensional space (which requires access to more function values since it leads to an increase in m'). This corresponds to using a relaxation $\widetilde{\mathcal{K}}_\varphi$ such that $\mathcal{K}_\varphi \subset \widetilde{\mathcal{K}}_\varphi \subset \widehat{\mathcal{K}}_\varphi$, for which $\sup_{\Sigma \in \widetilde{\mathcal{K}}_\varphi} \inf_{\Sigma' \in \mathcal{K}_\varphi} \|\Sigma - \Sigma'\|_{\mathrm{F}}$ is hopefully going to zero when the degree s goes to infinity, where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm. This is the case for several of the simple sets in the examples above, with a rate in $O(1/s^2)$, for hyperspheres [21], polynomials on $[-1, 1]^d$ [22], and trigonometric polynomials [11]. By increasing the degrees s until approximating the global optimum arbitrarily well, we obtain a "hierarchy" of optimization problems.

More precisely, this corresponds to replacing $\varphi(x) \in \mathbb{R}^m$ by $\tilde{\varphi}(x) = \binom{\varphi(x)}{\varphi^+(x)} \in \mathbb{R}^{\tilde{m}}$, and F by $\tilde{F} = \begin{pmatrix} F & 0 \\ 0 & 0 \end{pmatrix}$, with the function f defined by $f(x) = \varphi(x)^\top F \varphi(x) =$

$\tilde{\varphi}(x)^\top \tilde{F} \tilde{\varphi}(x)$. The convergence results in [11, 21, 22] correspond to the existence of $\varepsilon(\varphi, \varphi^+) > 0$ such that

$$\forall x \in \mathcal{X},\ f(x) \geq \varepsilon(\varphi, \varphi^+) \|F\|_{\mathrm{F}} \ \Rightarrow\ \exists \tilde{A} \succcurlyeq 0,\ \forall x \in \mathcal{X},\ f(x) = \tilde{\varphi}(x)^\top \tilde{A} \tilde{\varphi}(x).$$

The constant $\varepsilon(\varphi, \varphi^+)$ can be chosen as $O(1/s^2)$, where s is the degree of the polynomials defining $\tilde{\varphi}$.

Note that in practice, when using kernel formulations, using hierarchies simply means using a different kernel and more function evaluations.

3.6 Matrix-valued SOS

In Sect. 4, we will need to consider functions from $\mathcal{X}$ to some subspace $\mathcal{T}$ of $\mathbb{S}_p$ (the set of symmetric matrices of dimension p), and use characterizations of functions $f : \mathcal{X} \to \mathcal{T}$ that are linear in $\varphi(x)\varphi(x)^\top$ and such that for all $x \in \mathcal{X}$, $f(x) \succcurlyeq 0$. We assume the identiy matrix I belongs to $\mathcal{T}$.

This is an extension of the classical situation (where $p = 1$). We denote by $F \in \mathbb{R}^{mp \times mp}$ the linear form defined with blocks F_{ij} of size $m \times m$, for $i, j \in \{1, \dots, p\}$, such that

$$f(x) = F[\varphi(x)\varphi(x)^\top],$$

which is defined as $\forall x \in \mathcal{X},\ f(x)_{ij} = \varphi(x)^\top F_{ij} \varphi(x)$. The constraint that for all $x \in \mathcal{X}$, $f(x) \in \mathcal{T}$ is equivalent to $F \in \mathcal{V}_\varphi^\perp \otimes \mathbb{S}_p + \mathbb{S}_m \otimes \mathcal{T}$. Following [21, 23, 24], a sufficient condition for the matrix-non-negativity of f is $F \succcurlyeq 0$.

The condition $F \succcurlyeq 0$ is also necessary for some special cases. Indeed, if $\mathcal{T}$ is the set of diagonal matrices, we are then simply looking at p different non-negative polynomials, and if φ is such that we have a tight scalar SOS representation of non-negative functions, the condition is indeed necessary.

For the cases where we had the tightness guarantees in Sect. 3.5, it turns out that we have similar tightness guarantees, that is, if f is a degree $2r$ matrix-valued polynomials. We consider $\tilde{\varphi} = \binom{\varphi}{\varphi^+}$ leading to polynomials of degree $2s$, then if f has strictly positive-semidefinite values (that is, all eigenvalues greater than ε times some norm of f), then f is a matrix-SOS polynomial of degree $2s$. The constant ε can be taken as $O(1/s^2)$.

Indeed, all the proofs for hyperspheres [21], polynomials on $[-1, 1]^d$ [22], and trigonometric polynomials [11] are based on the same integral operator idea from [21] who showed how to extend it to the matrix domain. See a precise instance of such a result for trigonometric polynomials in Appendix A.

Link between matrix-SOS to tensor products In the min–max problem in Sect. 4, we will need to minimize quadratic forms in $\varphi(x)\varphi(x)^\top \in \mathbb{R}^{m \times m}$, rather than in $\varphi(x) \in \mathbb{R}^m$. Such a quadratic form is defined as $f(x) = \mathrm{tr}\big[F(\varphi(x)\varphi(x)^\top \otimes \varphi(x)\varphi(x)^\top)\big]$, and if the matrix-valued function $G : x \mapsto F[\varphi(x)\varphi(x)^\top]$ is such that $\forall x \in \mathcal{X},\ G(x) \succcurlyeq cI$, then $f(x) = \varphi(x)^\top G(x) \varphi(x) \geq c$ for all $x \in \mathcal{X}$. Thus the threshold for scalar-valued

quadratic forms in $\varphi(x)\varphi(x)^\top$ leads to (at least) the same threshold for matrix-valued quadratic forms in $\varphi(x)$.

4 SOS relaxations for min–max problems

We consider the min–max problem

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} g(x,y), \tag{15}$$

for a continuous function $g : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$ defined on the product of two compact sets $\mathcal{X}$ and $\mathcal{Y}$. We assume that we have two feature maps $\varphi : \mathcal{X} \to \mathbb{R}^m$ and $\psi : \mathcal{Y} \to \mathbb{R}^p$, such that $\|\varphi(x)\| = \|\psi(y)\| = 1$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, thus within the framework of Sect. 3. While the motivation is polynomials, this is not needed in most of this section.

We assume that the function g is a bilinear function of $\varphi\varphi^\top$ and $\psi\psi^\top$, that is, of the form

$$g(x,y) = \operatorname{tr}\Big[G\big(\psi(y)\psi(y)^\top \otimes \varphi(x)\varphi(x)^\top\big)\Big], \tag{16}$$

for a symmetrix matrix $G \in \mathbb{R}^{mp\times mp}$. By definition of the Kronecker product [25], we can see G as matrix defined by blocks G_{ij} of size $m \times m$, for $i, j \in \{1,\dots,p\}$, and Eq. (16) can be rewritten as:

$$g(x,y) = \sum_{i,j=1}^{p} \psi(y)_i\psi(y)_j \cdot \varphi(x)^\top G_{ij}\varphi(x).$$

Because of our unit norm assumptions for the feature maps, this includes linear forms in $\varphi\varphi^\top$ and $\psi\psi^\top$ (e.g., by considering all G_{ij} proportional to I, we obtain a linear form in $\psi\psi^\top$). For the examples in Sect. 3, such a representation exists for all multivariate polynomials in x and y.

The goal of this paper is to design SOS methods for this problem. Note that they will sometimes not be relaxations per se, as their values will not always be lower bounds on optimal values.

Special case of finite sets $\mathcal{Y}$. Throughout the paper, we will consider the special case of finite sets $\mathcal{Y}$, that is, the minimization of the maximum of finitely many polynomials, as it makes notations easier and sometimes allows further connections.

Notations Following Sect. 3, we denote by $\mathcal{V}_\varphi \subset \mathbb{R}^{m\times m}$ the span of all $\varphi(x)\varphi(x)^\top$, and $\mathcal{K}_\varphi \subset \mathbb{R}^{m\times m}$ the closure of its convex hull, with similar notations for $\mathcal{V}_\psi \subset \mathbb{R}^{p\times p}$ and $\mathcal{K}_\psi \subset \mathbb{R}^{p\times p}$.

The natural SOS formulation is to replace $\mathcal{K}_\varphi$ by its outer approximation $\widehat{\mathcal{K}}_\varphi = \{S \in \mathcal{V}_\varphi,\ S \succcurlyeq 0,\ \operatorname{tr}(S) = 1\} \supset \mathcal{K}_\varphi$ based on the affine hull of all $\varphi(x)\varphi(x)^\top$ on top of the positivity constraint, and, similarly, $\mathcal{K}_\psi$ by its outer approximation $\widehat{\mathcal{K}}_\psi = \{T \in \mathcal{V}_\psi,\ T \succcurlyeq 0,\ \operatorname{tr}(T) = 1\} \supset \mathcal{K}_\psi$. When using hierarchies, we may use

tighter sets $\widetilde{\mathcal{K}}_\varphi$ and $\widetilde{\mathcal{K}}_\psi$, which often corresponds to embedding φ in a bigger feature map.

We will also need $\mathcal{K}_{\varphi\otimes\varphi} \in \mathbb{R}^{m^2\times m^2}$ corresponding to the hull of all $\varphi(x)^{\otimes 4} = \varphi(x)\varphi(x)^\top \otimes \varphi(x)\varphi(x)^\top$, $x \in \mathcal{X}$, as well as it outer approximation $\widehat{\mathcal{K}}_{\varphi\otimes\varphi}$.

4.1 Existing SOS relaxation

The method of [1], which applies more generally (in particular to sets $\mathcal{Y}$ which are not simple), corresponds to an SOS formulation for Eq. (10), for a fixed probability measure $\mu \in \mathcal{P}(\mathcal{X})$ with full support, that is,

$$\min_{a:\mathcal{X}\to\mathbb{R}} \int_{\mathcal{X}} a(x)d\mu(x) \;\text{ such that }\; \forall (x,y) \in \mathcal{X}\times\mathcal{Y},\; a(x) \geq g(x,y),$$

and then the minimization of the resulting function a.

It can be cast as follows with our notations. In a first stage, assuming that one can compute $\Sigma = \mathbb{E}[\varphi(x)\varphi(x)^\top]$ for a distribution with full support on $\mathcal{X}$ (typically the uniform distribution), we solve

$$\begin{aligned}\min_{A\in\mathbb{R}^{m\times m},\; C\in\mathbb{R}^{mp\times mp}} &\operatorname{tr}(A\Sigma)\\ &\text{such that } C \succcurlyeq 0 \text{ and } \forall (x,y)\in\mathcal{X}\times\mathcal{Y},\; g(x,y)\\ &= \varphi(x)^\top A\varphi(x) - \operatorname{tr}\big[C\big(\psi(y)\psi(y)^\top \otimes \varphi(x)\varphi(x)^\top\big)\big],\end{aligned} \tag{17}$$

which approximates, with an SOS approach, the polynomial in x with the smallest expectation, which is above $g(x,y)$ for all $(x,y) \in \mathcal{X}\times\mathcal{Y}$. In the second stage, this polynomial a defined by the matrix A is minimized with an SOS method. This will converge when the degree of a is allowed to increase but requires approximating a non-polynomial function by a polynomial function, which may require a large degree. Moreover, it is typically not finitely convergent. Note that we could also minimize with respect to Σ in Eq. (17), but this leads to a non-convex problem. A natural algorithm for this non-convex problem is to perform alternating optimization, alternating between optimizing with respect to A and Σ, which improves the result but is not globally convergent in general (see Appendix B for more details). Finally, if the polynomial defined by A is minimized exactly, we obtain an upper-bound on the actual optimal value of Eq. (15).

Kernelization Using notations from Sect. 3, we can formulate the problem in Eq. (17) above as

$$\min_{A\in\mathbb{R}^{m\times m},\; C\in\mathbb{R}^{mp\times mp}} \operatorname{tr}(A\Sigma) \text{ such that } C \succcurlyeq 0 \text{ and } G - I\otimes A + C \in (\mathcal{V}_\varphi \otimes \mathcal{V}_\psi)^\perp$$

by definition of the vector space $\mathcal{V}_\varphi \otimes \mathcal{V}_\psi$. We can then introduce a Lagrange multiplier $M \in \mathcal{V}_\varphi \otimes \mathcal{V}_\psi$ to obtain by convex duality:

$$\max_{M\in\mathcal{V}_\varphi\otimes\mathcal{V}_\psi}\; \min_{A\in\mathbb{R}^{m\times m},\; C\succcurlyeq 0} \operatorname{tr}(A\Sigma) + \operatorname{tr}\big[M(G - I\otimes A + C)\big].$$

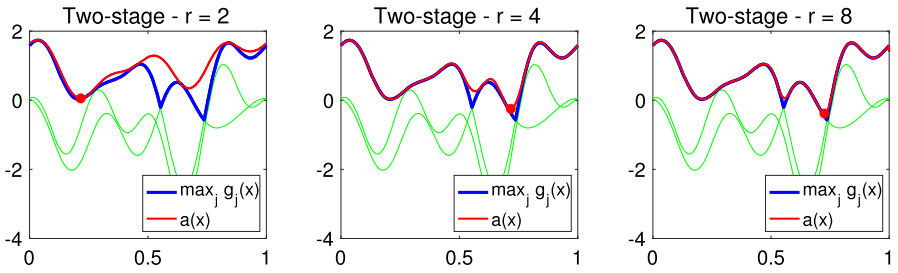

Fig. 1 Two-stage approach [1] for trigonometric polynomials in one dimension: three polynomials of maximal degree 2 on [0, 1] (in green), with their maximum (in blue), and the upper-bounding polynomial (in red), when using polynomials of degree r, with $r = 2$ (left), $r = 4$ (middle), and $r = 8$ (right)

We can then optimize with respect to $C \succcurlyeq 0$, which leads to the constraint $M \succcurlyeq 0$, and with respect to A, which leads to a linear constraint, that is:

$$\max_{M \in \mathcal{V}_\varphi \otimes \mathcal{V}_\psi} \operatorname{tr}(MG) \text{ such that } M \succcurlyeq 0 \text{ and } \widetilde{\operatorname{tr}}[M] = \Sigma,$$

where $\widetilde{\operatorname{tr}}[M] \in \mathbb{R}^{m \times m}$ denotes the "partial trace" defined as $\operatorname{tr}(N\widetilde{\operatorname{tr}}[M]) = \operatorname{tr}(M(N \otimes I))$ for any matrix $N \in \mathbb{R}^{m \times m}$, and $\mathbb{S}_p$ denotes the set of symmetric matrices of size p. In other words, $(\widetilde{\operatorname{tr}}[M])_{ij} = \operatorname{tr}(M_{ij})$.

It can be kernelized like in Sect. 3.3, in particular in situations where $\Sigma = \frac{1}{m} I$, which is the case when using a uniform distribution and φ obtained from orthonormal bases. Like in Eq. (14), we can then represent Σ as $\Sigma = \sum_{i=1}^{m'} \mu_i \varphi(x_i)\varphi(x_i)^\top$ for some $\mu \in \mathbb{R}^{m'}$. We thus solve

$$\max_{\alpha \in \mathbb{R}^{m' \times p'}} \sum_{i,j} \alpha_{ij} g(x_i, y_j) \text{ such that } \forall i, \sum_j \alpha_{ij} = \mu_i \text{ and}$$
$$\sum_{i,j} \alpha_{ij} \varphi(x_i)\varphi(x_i)^\top \otimes \psi(y_j)\psi(y_j)^\top \succcurlyeq 0, \tag{18}$$

and we recover $\varphi(x_i)^\top A \varphi(x_i)$ from the Lagrange multiplier for the constraint $\sum_{j=1}^{p'} \alpha_{ij} = \mu_i$. This is sufficient to minimize $a(x) = \varphi(x)^\top A \varphi(x)$ with an SOS method like described in Sect. 3.

Illustration We consider $\mathcal{X} = [0, 1]$ and trigonometric polynomials, with $\mathcal{Y} = \{1, 2, 3\}$. Thus, we aim to minimize the maximum of three trigonometric polynomials, which we take to have a maximal degree of 2. This is illustrated in Fig. 1, where we plot the three polynomials and the upper-bounding polynomial when using polynomials of degree r, with $r = 2, 4, 8$, where we can see the slow and in general only asymptotic convergence.

Alternative formulation for finite sets $\mathcal{Y}$. If $\mathcal{Y} = \{1, \ldots, p\}$, then our goal is to minimize the maximum of p multivariate polynomials $g_1, \ldots, g_p : \mathcal{X} \to \mathbb{R}$, and the simple one-stage formulation from [8, Section 3] (which applies more generally to finitely many rational functions) is based on the exact reformulation:

$$\min_{a \in \mathbb{R},\ x \in \mathcal{X}} a \ \text{ such that } \ \forall j \in \{1, \dots, p\},\ a - g_j(x) \geq 0,$$

which can be turned into an SOS formulation using standard constrained formulations (see [4]). This then leads to an upper-bound of the optimal value and often achieves a global minimizer, as illustrated in Sect. 5. However, this formulation does not extend to generic sets $\mathcal{Y}$.

4.2 Primal-dual SOS relaxation

We consider the following "exact" reformulation already presented in Eq. (4):

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} g(x, y) = \min_{\mu \in \mathcal{P}(\mathcal{X})} \max_{\nu : \mathcal{X} \to \mathcal{P}(\mathcal{Y})} \int_{\mathcal{X}} \int_{\mathcal{Y}} g(x, y) d\nu(y|x) d\mu(x),$$

where the maximization problem is replaced by the maximization of an expectation. Using the expression of g in Eq. (16), we can then use the bi-linearity of g and write the equation above as

$$\min_{\mu \in \mathcal{P}(\mathcal{X})} \max_{\nu : \mathcal{X} \to \mathcal{P}(\mathcal{Y})} \int_{\mathcal{X}} \operatorname{tr}\left[G\left(\int_{\mathcal{Y}} \psi(y)\psi(y)^\top d\nu(y|x) \otimes \varphi(x)\varphi(x)^\top \right) \right] d\mu(x),$$

and thus as (with no approximation yet), with $V(x) = \displaystyle\int_{\mathcal{Y}} \psi(y)\psi(y)^\top d\nu(y|x) \in \mathcal{K}_\psi$ (the hull of all $\psi(y)\psi(y)^\top$ for $y \in \mathcal{Y}$):

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} g(x, y) = \min_{\mu \in \mathcal{P}(\mathcal{X})} \max_{V : \mathcal{X} \to \mathcal{K}_\psi} \int_{\mathcal{X}} \operatorname{tr}\big[G\big(V(x) \otimes \varphi(x)\varphi(x)^\top \big) \big] d\mu(x). \tag{19}$$

Introducing the conditional distribution $\nu(\cdot|x)$ for all $x \in \mathcal{X}$, and the resulting pseudo-moment function $V : \mathcal{X} \to \mathcal{K}_\psi$ is key to avoiding the two-stage approach of [1].

We will now make a sequence of three relaxations to approximate the problem in Eq. (19) above.

Replacing $\mathcal{K}_\psi$ by $\widehat{\mathcal{K}}_\psi$ We first consider functions V with values in $\widehat{\mathcal{K}}_\psi$ (which is computationally more manageable) instead of $\mathcal{K}_\psi$ (which may not), leading to

$$\min_{\mu \in \mathcal{P}(\mathcal{X})} \max_{V : \mathcal{X} \to \widehat{\mathcal{K}}_\psi} \int_{\mathcal{X}} \operatorname{tr}\big[G\big(V(x) \otimes \varphi(x)\varphi(x)^\top \big) \big] d\mu(x), \tag{20}$$

which is always greater or equal to the optimal value of Eq. (19).

Parameterizing V by a matrix sum-of-squares The set $\widehat{\mathcal{K}}_\psi = \{T \succcurlyeq 0,\ \operatorname{tr}(T) = 1, T \in \mathcal{V}_\psi\}$ has a PSD constraint. Thus, as presented in Sect. 3.6, following [21, 23, 24], we can try to approximate it by a positive linear form in $\varphi(x)\varphi(x)^\top$ as

$$V(x) = T[\varphi(x)\varphi(x)^\top],$$

with $T \in \mathbb{R}^{mp\times mp}$ such that $T \succcurlyeq 0$, where, for $M \in \mathbb{R}^{m\times m}$, $T[M]$ denotes the symmetric matrix in $\mathbb{R}^{p\times p}$ such that for any symmetric matrix $N \in \mathbb{R}^{p\times p}$, $\mathrm{tr}\big(T[M]N\big) = \mathrm{tr}\big[T(N\otimes M)\big]$. In other words, if T is defined by blocks $T_{ij} \in \mathbb{R}^{m\times m}$ for $i, j \in \{1,\dots,p\}$, then $T[\varphi(x)\varphi(x)^\top]_{ij} = \varphi(x)^\top T_{ij}\varphi(x)$.

In order to impose that for all $x \in \mathcal{X}$, $V(x) \in \mathcal{V}_\psi$ and $\mathrm{tr}[V(x)] = 1$, we add the additional affine constraints

$$T \in \mathcal{V}_\varphi^\perp \otimes \mathbb{S}_p + \mathbb{S}_m \otimes \mathcal{V}_\psi, \quad \widetilde{\mathrm{tr}}[T] - I \in \mathcal{V}_\varphi^\perp,$$

where $\widetilde{\mathrm{tr}}[T] \in \mathbb{R}^{m\times m}$ denotes the "partial trace" defined as $\mathrm{tr}(M\widetilde{\mathrm{tr}}[T]) = \mathrm{tr}(T(M\otimes I))$ for any matrix $M \in \mathbb{R}^{m\times m}$, and $\mathbb{S}_p$ denotes the set of symmetric matrices of size p. In other words, $(\widetilde{\mathrm{tr}}[T])_{ij} = \mathrm{tr}(T_{ij})$. See Sect. 4.6 for an instantiation for discrete sets $\mathcal{Y}$, where notations slightly simplify.

We then obtain a problem where the measure μ only appears through the moment Σ of $\varphi(x)^{\otimes 4} = \varphi(x)\varphi(x)^\top \otimes \varphi(x)\varphi(x)^\top \in \mathbb{R}^{m^2\times m^2}$, since we have, using properties of Kronecker products:

$$\begin{aligned}\int_{\mathcal{X}} \mathrm{tr}\big[G\big(V(x)\otimes\varphi(x)\varphi(x)^\top\big)\big]d\mu(x) &= \sum_{i,j=1}^{p}\int_{\mathcal{X}} \varphi(x)^\top G_{ij}\varphi(x)\varphi(x)^\top T_{ij}\varphi(x)d\mu(x)\\ &= \mathrm{tr}\left(\Sigma\sum_{i,j=1}^{p} G_{ij}\otimes T_{ij}\right).\end{aligned}$$

We thus get a partially relaxed formulation, which cannot be solved yet by a semi-definite program (SDP), because of the set $\mathcal{K}_{\varphi\otimes\varphi}$:

$$\begin{aligned}\min_{\Sigma\in\mathbb{R}^{m^2\times m^2}} \max_{T\in\mathbb{R}^{mp\times mp}} \mathrm{tr}\left(\Sigma\sum_{i,j=1}^{p} G_{ij}\otimes T_{ij}\right) \text{ such that } \Sigma \in \mathcal{K}_{\varphi\otimes\varphi}\\ T \succcurlyeq 0,\ T \in \mathcal{V}_\varphi^\perp\otimes\mathbb{S}_p + \mathbb{S}_m\otimes\mathcal{V}_\psi,\ \widetilde{\mathrm{tr}}[T] - I \in \mathcal{V}_\varphi^\perp. \end{aligned} \tag{21}$$

Replacing $\mathcal{K}_{\varphi\otimes\varphi}$ by $\widehat{\mathcal{K}}_{\varphi\otimes\varphi}$ We obtain our final formulation, which can be solved as an SDP, where $\mathcal{K}_{\varphi\otimes\varphi}$ is replaced by $\widehat{\mathcal{K}}_{\varphi\otimes\varphi}$:

$$\begin{aligned}\min_{\Sigma\in\mathbb{R}^{m^2\times m^2}} \max_{T\in\mathbb{R}^{mp\times mp}} \mathrm{tr}\left(\Sigma\sum_{i,j=1}^{p} G_{ij}\otimes T_{ij}\right) \text{ such that } \Sigma \succcurlyeq 0,\ \Sigma \in \mathcal{V}_{\varphi\otimes\varphi},\ \mathrm{tr}(\Sigma) = 1\\ T \succcurlyeq 0,\ T \in \mathcal{V}_\varphi^\perp\otimes\mathbb{S}_p + \mathbb{S}_m\otimes\mathcal{V}_\psi,\ \widetilde{\mathrm{tr}}[T] - I \in \mathcal{V}_\varphi^\perp. \end{aligned} \tag{22}$$

We thus obtain a convex-concave min–max problem corresponding exactly to Eq. (9) in Sect. 2.2.

Alternative formulations We can then choose to transform it into a minimization problem akin to Eq. (7) by adding a Lagrange multiplier C for the constraint $T \succcurlyeq 0$, and $\Lambda \in \mathcal{V}_\varphi$ for $\widetilde{\mathrm{tr}}[T] - I \in \mathcal{V}_\varphi^\perp$, leading to:

$$\min_{\Sigma \in \mathbb{R}^{m^2 \times m^2},\ C \in \mathbb{R}^{mp \times mp},\ \Lambda \in \mathbb{R}^{m \times m}} \operatorname{tr}[\Lambda] \ \text{ such that } \ \Sigma \square G + C - \Lambda \otimes I \in \mathcal{V}_\varphi \otimes \mathcal{V}_\psi^\perp$$
$$\Sigma \in \mathcal{V}_{\varphi\otimes\varphi},\ \Sigma \succcurlyeq 0,\ \operatorname{tr}(\Sigma) = 1$$
$$\Lambda \in \mathcal{V}_\varphi,\ C \succcurlyeq 0, \tag{23}$$

where $\Sigma \square G \in \mathbb{R}^{mp \times mp}$ is defined by block as: $[\Sigma \square G]_{ij} = \Sigma_{ij} G_{ij} \in \mathbb{R}^{m \times m}$. This is the formulation used for solving the optimization problem empirically in Sect. 5.

Alternatively, we obtain a maximization problem akin to Eq. (8) from Eq. (22), by adding a Lagrange multiplier $A \succcurlyeq 0$ for the constraint $\Sigma \succcurlyeq 0$, and $c \in \mathbb{R}$ for the constraint $\operatorname{tr}\Sigma = 1$:

$$\max_{T \in \mathbb{R}^{mp \times mp},\ A \in \mathbb{R}^{m^2 \times p^2},\ c \in \mathbb{R}} c \ \text{ such that } \ T \circ G - cI - A \in \mathcal{V}_{\varphi\otimes\varphi}^\perp$$
$$T \succcurlyeq 0,\ T \in \mathcal{V}_\varphi^\perp \otimes \mathbb{S}_p + \mathbb{S}_m \otimes \mathcal{V}_\psi,\ \widetilde{\operatorname{tr}}[T] - I \in \mathcal{V}_\varphi^\top$$
$$A \succcurlyeq 0. \tag{24}$$

This formulation will be used in the convergence proof in Sect. 4.5. It is implementable as a semidefinite program as soon as the sets $\mathcal{V}_\psi$, $\mathcal{V}_\varphi$, and $\mathcal{V}_{\varphi\otimes\varphi}$ can be represented by finitely many linear constraints.

Summary Overall, the formulation is obtained through 3 approximations:

- Replacing $\mathcal{K}_\psi$ by $\widehat{\mathcal{K}}_\psi$. If this approximation is exact, then the maximization in y is exact, and we obtain lower bounds. This is, for example the case for $\mathcal{Y} = \{1, \dots, p\}$, and also for degree 2 polynomials. Otherwise, the equal values of problems in Eqs. (22), (23), and (24) may be above or below the optimal value.
- Parameterizing all functions $V : \mathcal{X} \to \mathcal{V}_\psi \cap \mathbb{S}_p^+$ by a matrix sum-of-squares [21, 23, 24]. This can only be exact if the function V is a polynomial, with the extra approximation due to the potential non-tightness of matrix SOS.
- Replacing $\mathcal{K}_{\varphi\otimes\varphi}$ by $\widehat{\mathcal{K}}_{\varphi\otimes\varphi}$. This is a typical SOS relaxation problem.

These approximations are discussed in Sect. 4.5.

4.3 Kernelization

We assume that we have m points $x_1, \dots, x_m$ such that the corresponding kernel matrix is invertible, complemented by $m' - m$ points $x_{m+1}, \dots, x_{m'}$ such that $\mathcal{V}_\varphi$ is spanned by $\varphi(x_1)\varphi(x_1)^\top, \dots, \varphi(x_{m'})\varphi(x_{m'})^\top$, and finally $m'' - m'$ points such that $\mathcal{V}_{\varphi\otimes\varphi}$ is spanned by $\varphi(x_1)^{\otimes 4}, \dots, \varphi(x_{m''})^{\otimes 4}$. We denote by $K' \in \mathbb{R}^{m' \times m'}$ the kernel matrix of the first m' points. The matrix K' is not invertible, but $K' \circ K'$ (with $\circ$ the element-wise product) is, because $\varphi(x_1)\varphi(x_1)^\top, \dots, \varphi(x_{m'})\varphi(x_{m'})^\top$ is a basis of $\mathcal{V}$. We denote by $K'' \in \mathbb{R}^{m' \times m''}$ the kernel matrix between the m' first points and all m'' points.

We can then express for all $i \in \{1, \dots, m''\}$, $\varphi(x_i)\varphi(x_i)^\top = \sum_{j=1}^{m'} N_{ji}\varphi(x_j)\varphi(x_j)^\top$, with the matrix $N \in \mathbb{R}^{m' \times m''}$ equal to $N = (K' \circ K')^{-1} K''$.

We can write Eq. (23) with $\Sigma = \sum_{i=1}^{m''} \alpha_i \varphi(x_i)^{\otimes 4}$, $C = \sum_{i=1}^{m'} D_i \otimes \varphi(x_i)\varphi(x_i)^\top$, and $\Lambda = \sum_{i=1}^{m'} \lambda_i \varphi(x_i)\varphi(x_i)^\top$, and get the optimization problem (which is an SDP which we use in our experiments Sect. 5):

$$\begin{aligned} &\min_{\alpha \in \mathbb{R}^{m''},\ \lambda \in \mathbb{R}^{m'},\ D_1, \ldots, D_{m'} \in \mathbb{R}^{p' \times p'}} \sum_{i=1}^{m'} \lambda_i \\ \text{such that } &\forall i \in \{1, \ldots, m'\},\ j \in \{1, \ldots, p'\},\ \psi(y_j)^\top D_i \psi(y_j) - \lambda_i + [N \operatorname{Diag}(\alpha) G]_{ij} = 0 \\ &\sum_{i=1}^{m'} D_i \otimes \varphi(x_i)\varphi(x_i)^\top \succcurlyeq 0,\ \sum_{i=1}^{m''} \alpha_i = 1,\ \sum_{i=1}^{m''} \alpha_i \varphi(x_i)^{\otimes 4} \succcurlyeq 0. \end{aligned}$$

From the vector α, we can obtain a potential minimizer using algorithms from Sect. 3.4, with the possibility of full kernelization as in Sect. 3.3, where $G \in \mathbb{R}^{m'' \times p'}$ is the matrix of evaluations $g(x_i, y_j)$.

4.4 A posteriori guarantees

Since we have used relaxations of both maximization and minimization problems, we do not obtain, in general, an upper or lower bound, except in some special cases that we now describe.

If the feature map ψ is such that $\widehat{\mathcal{K}}_\psi = \mathcal{K}_\psi$, then from the matrix $T \in \mathbb{R}^{mp \times mp}$, we get $V : \mathcal{X} \to \mathcal{K}_\psi$, and thus a feasible dual point for Eq. (19). Therefore the value of the SOS formulation is always below the true one. If Σ is represented by a singleton $\varphi(x_*)\varphi(x_*)^\top$, then if $V(x_*)$ is such that $\max_{y \in \mathcal{Y}} L(x_*, y) = \operatorname{tr}[M(\varphi(x_*)\varphi(x_*)^\top \otimes V(x_*))]$, we have a tight solution. This happens in our simulations.

If $\widehat{\mathcal{K}}_\psi \supsetneq \mathcal{K}_\psi$, then, when Σ is represented by a singleton $\varphi(x_*)\varphi(x_*)^\top$, if $V(x_*)$ is such that $\max_{y \in \mathcal{Y}} L(x_*, y) = \operatorname{tr}[M(\varphi(x_*)\varphi(x_*)^\top \otimes V(x_*))]$, we only know that we have an upper-bound on the true value.

4.5 A priori guarantees

In this section, we focus primarily on the situation where $\widehat{\mathcal{K}}_\psi = \mathcal{K}_\psi$, so we do not have to use hierarchies on $\mathcal{Y}$. If this is not the case, we can use another relaxation $\widetilde{K}_\psi$ that would lead to an extra approximation factor that goes to zero as the degree of the hierarchy on y goes to infinity. Still, the precise details are out of the scope of this paper.

The main new result shows that for the polynomial examples in Sect. 3.1, the partially relaxed problem in Eq. (21) can be solved through hierarchies with arbitrary precisions. We make the following assumptions:

(A1) $\widehat{\mathcal{K}}_\psi = \mathcal{K}_\psi$, so that our relaxation is a lower-bound.

(A2) Given a one-dimensional Lipschitz-continuous function $g : \mathcal{X} \to \mathbb{R}$, it can be approximated by a quadratic form in $\binom{\varphi}{\varphi_1^+}$, where φ_1^+ includes additional monomials, and we denote by $\varepsilon^{\text{app}}(\varphi, \varphi_1^+)$ the approximation constant so that

for all g, there exists a quadratic form in $\tilde{\varphi}$ defined by the matrix $\tilde{H}$, such that

$$\forall x \in \mathcal{X},\ |g(x) - \tilde{\varphi}(x)^\top \tilde{H} \tilde{\varphi}(x)| \leq \mathrm{Lip}(g) \cdot \varepsilon^{\mathrm{app}}(\varphi, \varphi_1^+).$$

It is known that, given φ, we can make $\varepsilon^{\mathrm{app}}(\varphi, \varphi_1^+)$ as small as desired by increasing the degree of the polynomials, with well-studied convergence rates, through "Jackson's inequalities" [26].

(A3) We solve the equivalent optimization problems in Eqs. (22), (23), or (24) with φ replaced by $\tilde{\varphi} = \binom{\varphi}{\varphi^+}$, where $\varphi^+ = \binom{\varphi_1^+}{\varphi_2^+}$ includes additional mononials on top of φ_1^+. We denote by $\varepsilon^{\mathrm{SOS}}(\varphi, \varphi_1^+, \varphi_2^+)$ the SOS approximability ratio defined in Sect. 3.5 as, for any $H \in \mathbb{R}^{(m+m_1)s \times (m+m_1)s}$:

$$\begin{aligned} &\forall x \in \mathcal{X},\ H[\varphi(x)\varphi(x)^\top] \succcurlyeq \varepsilon^{\mathrm{SOS}}(\varphi, \varphi_1^+, \varphi_2^+) \|H\|_{\mathrm{F}} I \\ &\quad \Rightarrow\ \exists \tilde{A} \succcurlyeq 0,\ \forall x \in \mathcal{X},\ H[\varphi(x)\varphi(x)^\top] = \tilde{A}[\tilde{\varphi}(x)\tilde{\varphi}(x)^\top]. \end{aligned}$$

We select the threshold to have a similar result for scalar-valued quadratic forms in $\varphi(x)\varphi(x)^\top$, as described at the end of Sect. 3.6. We know from Sect. 3.6 that, given φ and φ_1^+, we can make $\varepsilon^{\mathrm{SOS}}(\varphi, \varphi_1^+, \varphi_2^+)$ as small as desired by increasing the degree of the polynomials.

Note that we only need to divide φ^+ in $\binom{\varphi_1^+}{\varphi_2^+}$ for the proof, as the algorithm is oblivious to this distinction. Our main result follows.

Theorem 1 *Let G be defined by a polynomial as in Eq. (16) and $\varepsilon > 0$. Assume (A1), (A2), and (A3). Then there exist feature maps φ_1^+ and φ_2^+ such that the optimal value of the SOS primal-dual relaxation is within ε of the optimal value.*

Proof This requires obtaining SOS polynomial approximations to the matrix-valued function V obtained in Eq. (20). To obtain a finite convergence, we would need to represent one of the many optimal V's exactly. Here we will consider a specific approximation based on Von Neumann entropy regularization and start with a smoothing lemma.

Lemma 1 *Let $B \in \mathbb{S}_p$ and $\eta > 0$. Let $W_\eta(B)$ be the unique maximizer of $\mathrm{tr}[BW] - \eta \mathrm{tr}[W \log W]$ such that $W \succcurlyeq 0$, $\mathrm{tr}(W) = 1$, and $W \in \mathcal{V}_\psi \subset \mathbb{R}^p$. Then W_η is a $(1/\eta)$-Lipschitz-continuous function of B, and*

$$0 \leq \max_{W \succcurlyeq 0,\ \mathrm{tr}(W)=1,\ W \in \mathcal{V}_\psi} \mathrm{tr}[BW] - \mathrm{tr}[BW_\eta(B)] \leq \eta \log p.$$

Proof Since the function $W \mapsto \mathrm{tr}[W \log W]$ is 1-strongly convex with respect to the nuclear norm on the set $\{W \succcurlyeq 0, \mathrm{tr}(W) = 1\}$ [27], W_η is such that $\|W_\eta(B) - W_\eta(B')\|_* \leq \frac{1}{\eta}\|B - B'\|_{\mathrm{op}}$ [28], where $\|\cdot\|_*$ denotes the nuclear norm. The bound is obtained by looking at eigenvalues of V and using classical bound on entropies [29]. □

We can now build a feasible point for Eq. (24) which will lead to the desired bound. Following the discussion at the end of Sect. 2, there are many optimal candidates for $x \mapsto V(x)$. In this proof, we propose a dual candidate V based on maximizing approximately $g(x, y)$ for *all* $x \in \mathcal{X}$, and not only at the minimizer x_*. While it allows to show asymptotic convergence, it is not sufficient to show finite convergence. Interestingly, in our proof, we end up approximating $\max_{y\in\mathcal{Y}} g(x, y)$ by a polynomial in x with a specific form (see Eq. (25) below), like done by [1] (see Sect. 4.1 for the precise formulation). However, this is only used in the proof, and not within the algorithm. Finding a proof that circumvents this need for polynomial approximation would strenghten the result.

We consider $\eta > 0$, and the function $V : x \mapsto W_\eta\big(G[\varphi(x)\varphi(x)^\top]\big)$ obtained from Lemma 1. By construction, $\forall x \in \mathcal{X}, V(x) \in \widehat{\mathcal{K}}_\psi$, V is Lipschitz-continuous with constant proportional to $\|G\|_{\mathrm{F}}/\eta$ (with constants that depends on φ), since

$$\|V(x') - V(x)\|_* \leq \frac{1}{\eta}\|G'[\varphi(x)\varphi(x')^\top] - G[\varphi(x)\varphi(x)^\top]\|_{\mathrm{op}} \leq \frac{1}{\eta}\|G\|_{\mathrm{F}}\mathrm{Lip}(\varphi).$$

Moreover, from Lemma 1, we have

$$0 \leq \max_{y\in\mathcal{Y}} g(x, y) - \mathrm{tr}\big(V(x)G[\varphi(x)\varphi(x)^\top]\big) \leq \eta \log p. \tag{25}$$

We can then use Assumption *(A2)* and approximate V by a quadratic form in $\binom{\varphi}{\varphi_1^+}$. We thus find a matrix $U \in \mathbb{R}^{(m+m_1)p\times(m+m_1)p}$ such that all affine constraints on values of V are still satisfied, that is, $U \in \mathcal{V}_\varphi^\perp \otimes \mathbb{S}_p + \mathbb{S}_m \otimes \mathcal{V}_\psi$ and $\widetilde{\mathrm{tr}}[U] - I \in \mathcal{V}_\varphi^\top$, and

$$\forall x \in \mathcal{X},\ \|V(x) - U[\varphi(x)\varphi(x)^\top]\|_{\mathrm{op}} \leq C \cdot \frac{1}{\eta}\|G\|_{\mathrm{F}} \cdot \varepsilon^{(\mathrm{app})}(\varphi, \varphi_1^+)$$

for some constant C that is independent of G and φ_1^+. Thus for all $x \in \mathcal{X}$, $U[\varphi(x)\varphi(x)^\top] \succcurlyeq -C \cdot \frac{1}{\eta}\|G\|_{\mathrm{F}} \cdot \varepsilon^{\mathrm{app}}(\varphi, \varphi_1^+)I$, which implies from Assumption *(A3)* that

$$U[\varphi(x)\varphi(x)^\top] + \left[\varepsilon^{\mathrm{SOS}}(\varphi, \varphi_1^+, \varphi_2^+)\|U\|_{\mathrm{F}} + C \cdot \frac{1}{\eta}\|G\|_{\mathrm{F}} \cdot \varepsilon^{\mathrm{app}}(\varphi, \varphi_1^+)\right] I$$

is a sum-of-squares and satisfies all other affine constraints. We denote by T the corresponding matrix, which is feasible for Eq. (24). Moreover, because of Eq. (25), we have, for all $x \in \mathcal{X}$,

$$\begin{aligned}\mathrm{tr}\big(T[\varphi(x)\varphi(x)^\top]G[\varphi(x)\varphi(x)^\top]\big) &\geq \min_{x'\in\mathcal{X}} \max_{y\in\mathcal{Y}} g(x', y) \\ &-\eta \log p - \left[\varepsilon^{\mathrm{SOS}}(\varphi, \varphi_1^+, \varphi_2^+)\|U\|_{\mathrm{F}} + \frac{C}{\eta}\|G\|_{\mathrm{F}} \cdot \varepsilon^{\mathrm{app}}(\varphi, \varphi_1^+)\right],\end{aligned}$$

and thus, applying Assumption *(A3)* again, $T \circ G - \min_{x' \in \mathcal{X}} \max_{y \in \mathcal{Y}} g(x', y) - \eta \log p + \big[\varepsilon^{\rm SOS}(\varphi, \varphi_1^+, \varphi_2^+)(\|U\|_{\rm F} + \|T \circ G\|_{\rm F}) + \frac{C}{\eta}\|G\|_{\rm F} \cdot \varepsilon^{\rm app}(\varphi, \varphi_1^+)\big]$ is a sum of squares, and thus we obtain an approximation of $\min_{x' \in \mathcal{X}} \max_{y \in \mathcal{Y}} g(x', y)$ up to $\eta \log p + \varepsilon^{\rm SOS}(\varphi, \varphi_1^+, \varphi_2^+)(\|U\|_{\rm F} + \|T \circ G\|_{\rm F}) + \frac{C}{\eta}\|G\|_{\rm F} \cdot \varepsilon^{\rm app}(\varphi, \varphi_1^+)$. Now, given $\varepsilon > 0$, we take $\eta = \frac{1}{3 \log p}$, then select φ_1^+ such that $\frac{C}{\eta}\|G\|_{\rm F} \cdot \varepsilon^{\rm app}(\varphi, \varphi_1^+)$ is smaller than $\varepsilon/3$, and finally select φ_2^+ such that $\varepsilon^{\rm SOS}(\varphi, \varphi_1^+, \varphi_2^+)(\|U\|_{\rm F} + \|T \circ G\|_{\rm F})$ is less than $\varepsilon/3$. This leads to desired approximation within ε. □

We can make the following observations:

- The hierarchy often empirically converges in finitely many iterations, but we cannot find provable sufficient conditions. It would be interesting to see if, assuming that the polynomial is convex-concave, we could use tools from [30] to prove such convergence.
- To obtain a convergence rate, we would need to be able to characterize the dependence of $\varepsilon^{\rm SOS}(\varphi, \varphi_1^+, \varphi_2^+)$ on φ_1^+, which we leave for future work.

4.6 Special case $\mathcal{Y} = \{1, \dots, p\}$

This corresponds to having $\mathcal{K}_\psi = \widehat{\mathcal{K}}_\psi$ the set of PSD diagonal matrices with unit trace. We can then simplify notations and solve

$$\min_{x \in \mathcal{X}} \max_{j \in \{1, \dots, p\}} \varphi(x)^\top G_j \varphi(x),$$

with $G_1, \dots, G_p \in \mathbb{S}_m$. We then have V diagonal, with diagonal elements $v_j(x) = \varphi(x)^\top T_j \varphi(x)$, with $T_j \succcurlyeq 0$, and $\sum_{j=1}^p T_j - I \in \mathcal{V}_\varphi^\perp$. We thus obtain the min/max formulation corresponding to Eq. (22):

$$\begin{aligned} &\min_{\Sigma \in \mathbb{R}^{m^2 \times m^2}} \max_{T_1, \dots, T_p \in \in \mathbb{R}^{m \times m}} {\rm tr}\left[\Sigma \cdot \sum_{j=1}^p G_j \otimes T_j\right] \text{ such that } \Sigma \succcurlyeq 0, \\ &\Sigma \in \mathcal{V}_{\varphi \otimes \varphi},\ {\rm tr}(\Sigma) = 1 T_1, \dots, T_p \succcurlyeq 0,\ \sum_{j=1}^p T_j - I \in \mathcal{V}_\varphi^\perp. \end{aligned} \tag{26}$$

We also get the minimization formulation, which is the one used in experiments, corresponding to Eq. (23):

$$\begin{aligned} \min_{\Sigma \in \mathbb{R}^{m^2 \times m^2},\ \Lambda \in \mathbb{R}^{m \times m}} {\rm tr}[\Lambda] \text{ such that } &\forall j \in \{1, \dots, p\},\ \Sigma[G_j] \preccurlyeq \Lambda \\ &\Sigma \in \mathcal{V}_{\varphi \otimes \varphi},\ \Sigma \succcurlyeq 0,\ {\rm tr}(\Sigma) = 1 \\ &\Lambda \in \mathcal{V}_\varphi. \end{aligned} \tag{27}$$

We also get a maximization formulation, corresponding to Eq. (24), and leading to a nice interpretation below:

$$\max_{T_1,\ldots,T_p\in\mathbb{R}^{m\times m}} c \text{ such that } \sum_{j=1}^{p} G_j \otimes T_j - cI - A \in \mathcal{V}_{\varphi\otimes\varphi}^{\perp}$$
$$T_1,\ldots,T_p \succcurlyeq 0,\ \sum_{j=1}^{p} T_j - I \in \mathcal{V}_{\varphi}^{\perp}. \tag{28}$$

Kernelization Empirically, we solve

$$\min_{\alpha\in\mathbb{R}^{m''},\ \lambda\in\mathbb{R}^{m'}} \sum_{i=1}^{m'} \lambda_i \text{ such that } \forall j \in \{1,\ldots,p\},\ \sum_{i=1}^{m''} \alpha_i g_j(x_i)\varphi(x_i)\varphi(x_i)^\top \preccurlyeq$$
$$\sum_{i=1}^{m'} \lambda_i \varphi(x_i)\varphi(x_i)^\top \sum_{i=1}^{m''} \alpha_i = 1,\ \sum_{i=1}^{m''} \alpha_i \varphi(x_i)\varphi(x_i)^\top \otimes \varphi(x_i)\varphi(x_i)^\top \succcurlyeq 0.$$

We obtain the matrices $T_1,\ldots,T_p$ as Lagrange multipliers for the PSD constraints.

Relationship with Putinar's Positivstellensatz An interesting parallel with Putinar's Positivstellensatz [31] can be made. We consider p multi-variate polynomials $g_1,\ldots,g_p$, with $\mathcal{X}\subset\mathbb{R}^d$ one of the simple sets described in Sect. 3. Because of the approximation result in Sect. 4.5, we know that if $\min_{x\in\mathcal{X}} \max_{j=1,\ldots,p} g_j(x)$ is strictly positive, there is a level of the hierarchy of polynomials so that our relaxation also has strictly positive values, and, in fact, the converse is also true. Thus, using the maximization formulation from Eq. (28), $\min_{x\in\mathcal{X}} \max_{j=1,\ldots,p} g_j(x) > 0$, if and only if there exists $c > 0$ and sum-of-square (that is, PSD quadratic forms in φ) polynomials q_0 (represented by A), and $q_1,\ldots,q_p$, represented by $T_1,\ldots,T_p$, such that

$$\forall x \in \mathcal{X},\ c = \sum_{j=1}^{p} g_j(x)q_j(x) - q_0(x),$$

and such that $q_1(x) + \cdots + q_p(x) = 1$ for all $x \in \mathcal{X}$.

Without loss of the generality, we can take $c = 1$, and we have shown that

$$\min_{x\in\mathcal{X}} \max_{j=1,\ldots,p} g_j(x) > 0$$

if and only if there exist SOS polynomials (based on the feature vector φ) $q_0,\ldots,q_p$ such that

$$\forall x \in \mathcal{X},\ -1 = \sum_{j=1}^{p} (-g_j(x))q_j(x) + q_0(x)$$

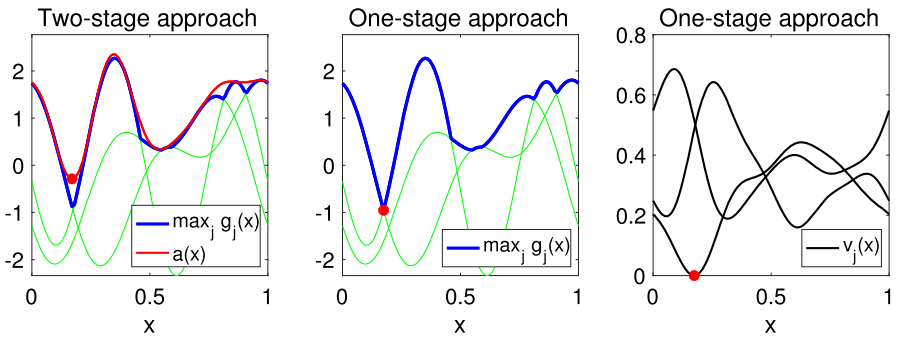

Fig. 2 Minimization of the maximum of 3 trigonometric polynomials on $[0, 1]$: two-stage approach of [1] (left), one-stage primal-dual approach (middle), functions v_j, $j = 1, 2, 3$ from the two-stage approach (right)

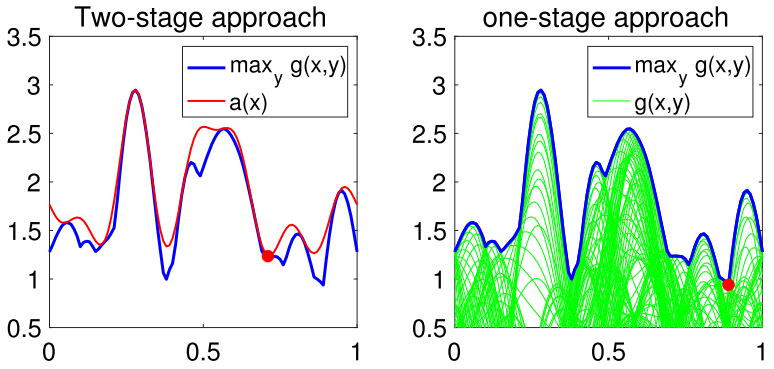

Fig. 3 Minimization of the maximum of a bivariate trigonometric polynomial with $\mathcal{X} = \mathcal{Y} = [0, 1]$. Two-stage approach of [1] (left), one-stage primal-dual approach (right)

and $q_1(x) + \cdots + q_p(x)$ is constant on $\mathcal{X}$.

Without the last constraint, this turns out to be exactly the Putinar certificate for the positivity of -1 on the set $\mathcal{A} = \{x \in \mathbb{R}^d, \ \forall j \in \{1, \dots, p\}, \ -g_j(x) \geq 0\}$, and thus a certificate for the emptiness of that set. Given that

$$\min_{x \in \mathcal{X}} \max_{j=1,\dots,p} g_j(x) \leq 0 \ \Leftrightarrow \ \exists x \in \mathcal{X}, \ \forall j \in \{1, \dots, p\}, \ g_j(x) \leq 0 \ \Leftrightarrow \ \mathcal{X} \cap \mathcal{A} \neq \varnothing,$$

we obtained a feasibility certificate similar to the one obtained for Putinar. Note that the original Putinar certificate does require an extra assumption, e.g., that one of the sets $\{x \in \mathbb{R}^d, \ -g_j(x) \geq 0\}$ is bounded.

Note moreover that with our assumptions from Sect. 3.1, SOS-polynomials that are PSD quadratic forms in φ have a slightly different meaning; that is, for example, for the unit Euclidean ball, they correspond to $(1 - \|x\|_2^2)u(x) + v(x)$, where u and v are regular sums-of-squares. This leads to the following proposition (where we have replaced g_j by $-g_j$ to match classical certificates, and we have dropped the constraint of summing to a constant, which is not necessary).

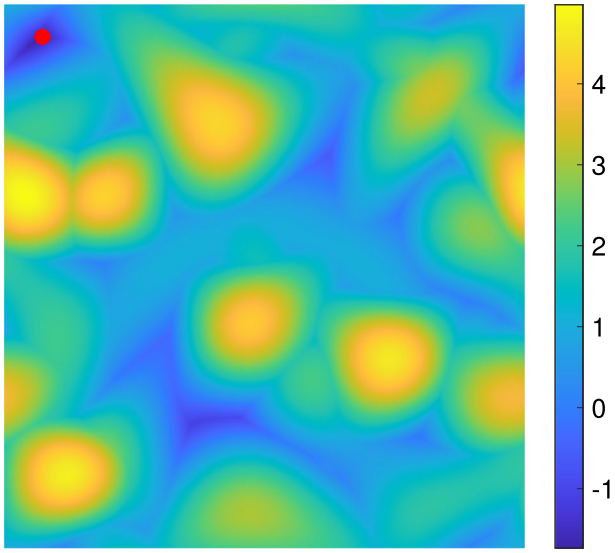

Fig. 4 Minimization of the maximum 4 trigonometric polynomials on $[0, 1]^2$, with the maximizer in red

Proposition 2 *Let $g_1, \dots, g_p$ be p multivariate polynomials on $\mathbb{R}^d$. Then, the set*

$$\{x \in \mathbb{R}^d,\ \|x\|_2^2 \leq 1,\ \forall j \in \{1, \dots, p\},\ g_j(x) \geq 0\}$$

is empty if and only if there exists polynomials $u_0, v_0, u_1, v_1, \dots, u_p, v_p$ that are sums-of-squares such that

$$\forall x \in \mathbb{R}^d,\ -1 = \sum_{j=1}^{p} g_j(x)\big[(1 - \|x\|_2^2)u_j(x) + v_j(x)\big] + (1 - \|x\|_2^2)u_0(x) + v_0(x).$$

It is weaker than Putinar's certificate, which would not need $u_1, \dots, u_p$. Still, it could be extended to continuous situations where the set $\mathcal{Y}$ (and the corresponding feature map ψ) in our min–max formulation leads to tight SOS formulations, for example, for polynomials in $[-1, 1]$.

5 Experiments

In this section, we provide illustrative experiments where we obtain tight relaxations on small problems. See https://www.di.ens.fr/~fbach/sos_min_max.zip for Matlab code reproducing these experiments.

Minimizing the maximum of univariate trigonometric polynomials See Fig. 2, where we obtain a tight relaxation with our one-stage approach (which always leads to lower bounds in this situation), while the two-stage approach from [1] does not. Here, the

one-stage approach from [8] (which always leads to upper bounds) can be applied and is also tight because SOS relaxations are tight in dimension one.

We also plot the optimal function $v_j(x) = \varphi(x)^\top T_j \varphi(x)$, $j \in \{1, \dots, p\}$, which are non-negative and sum to one, and, at x_*, have non-zero values only for the j's attaining the maximum in $\max_{j \in \{1,\dots,p\}} g_j(x_*)$.

Maximizing the maximum of bivariate trigonometric polynomials See Fig. 4 for an example with a tight relaxation, while the two-stage approach from [1] does not (here, the one-stage approach of [8] cannot be applied).

Min–max optimization of a trigonometric polynomial on $[0, 1]^2$. See Fig. 3 for an example with a tight relaxation.

6 Conclusion

In this paper, we proposed an SOS formulation for min–max problems over polynomials and provided a convergence proof when degrees of polynomials are allowed to increase. This work opens up several avenues for future work, such as (a) infinite-dimensional extensions for smooth functions by adding proper regularization like done by [16] for plain minimization, (b) finding sufficient conditions for either finite convergence or an explicit rate, and (c) exploring how the min–max approach relates to the several Positivstellensatz from the literature.

Acknowledgements The comments and suggestions of the anonymous reviewers were greatly appreciated.

Appendix A Convergence rates of matrix-valued SOS

We extend the proof of [11, Theorem 1] to matrix-valued polynomials, using the same technique as [21], and following the notations of [11] closely.

Proposition 3 *Let $r > 0$ and $s \geq 3r$, and $\varepsilon(s) = \left[\left(1 - \frac{6r^2}{s^2}\right)^{-d} - 1\right] \sim_{s \to +\infty} \frac{6r^2 d}{s^2}$. For any multivariate matrix-valued trigonometric polynomial f of degree less than $2r$, written $f(x) = \sum_{\|\omega\|_\infty \leq 2r} \hat{f}(\omega) e^{2i\pi\omega^\top x}$,*

$$\forall x \in [0,1]^d,\ f(x) \succcurlyeq \varepsilon(s) \sum_{\|\omega\|_\infty \leq 2r,\ \omega \neq 0} \|\hat{f}(\omega)\|_{\mathrm{op}}$$
$$\Rightarrow\ f \text{ is a sum of squares of polynomials of degree } s.$$

Proof We consider the following integral operator on 1-periodic matrix-valued functions on $[0, 1]^d$, defined as

$$Th(x) = \int_{[0,1]^d} |q(x - y)|^2 h(y) dy, \tag{A1}$$

for a well-chosen 1-periodic function q which is a trigonometric polynomial of degree s. The function $x \mapsto |q(x - y)|^2$ is an element of the finite-dimensional

cone of SOS polynomials of degree s, thus, by design, if h has positive semi-definite values, then Th is a sum of squares of matrix polynomials of degree less than s. We will find h such that $Th = f$.

In the Fourier domain, since convolutions lead to pointwise multiplication and vice-versa, we have for all $\omega \in \mathbb{Z}^d$, where $\hat{q} * \hat{q}(\omega)$ is a shorthand for $(\hat{q} * \hat{q})(\omega)$:

$$\widehat{Th}(\omega) = \hat{q} * \hat{q}(\omega) \cdot \hat{h}(\omega),$$

and thus, the candidate h is defined by its Fourier series, which is equal to zero for $\|\omega\|_\infty > 2r$, and to

$$\frac{\hat{f}(\omega)}{\hat{q} * \hat{q}(\omega)}$$

otherwise. If we impose that $\hat{q} * \hat{q}(0) = 1$, we then have

$$\begin{aligned} f - h &= \sum_{\omega \in \mathbb{Z}^d} \hat{f}(\omega)\Big(1 - \frac{1}{\hat{q} * \hat{q}(\omega)}\Big) \exp(2i\pi\omega^\top \cdot) \\ &= \sum_{\omega \neq 0} \hat{f}(\omega)\Big(1 - \frac{1}{\hat{q} * \hat{q}(\omega)}\Big) \exp(2i\pi\omega^\top \cdot). \end{aligned}$$

We then get:

$$\sup_{x \in [0,1]^d} \|f(x) - h(x)\|_{\text{op}} \leq \sum_{\omega \neq 0} \big\|\hat{f}(\omega)\big\|_{\text{op}} \cdot \max_{\|\omega\|_\infty \leq 2r} \Big|\frac{1}{\hat{q} * \hat{q}(\omega)} - 1\Big|. \tag{A2}$$

With the choice $\hat{q}(\omega) = a \prod_{i=1}^d \Big(1 - \frac{|\omega_i|}{s}\Big)_+$, with a a normalizing constant, we get $\hat{q} * \hat{q}(0) = 1$ and $\max_{\|\omega\|_\infty \leq 2r} \Big|\frac{1}{\hat{q} * \hat{q}(\omega)} - 1\Big| \leq \varepsilon(s)$ (see [11] for details). Thus, for all $x \in [0, 1]^d$, using Eq. (A2) and the assumption on f:

$$h(x) = f(x) - (f(x) - h(x)) \succcurlyeq \varepsilon(s) \sum_{\omega \neq 0} \|\hat{f}(\omega)\|_{\text{op}} - \varepsilon(s) \sum_{\omega \neq 0} \|\hat{f}(\omega)\|_{\text{op}} = 0,$$

which leads to the desired result. □

Appendix B Alternating optimization for the two-stage approach

In this section, we explore briefly the possibility evoked in Sect. 4.1 of trying to minimize Eq. (17) with respect to Σ as well. This is a non-convex problem, and alternating optimization has a particularly simple formulation. Indeed, in the kernelized version in Eq. (18), this corresponds to replacing μ by the previous value of α and iterating. Since the first upper-bound is minimized exactly, at the second iteration and all later ones,

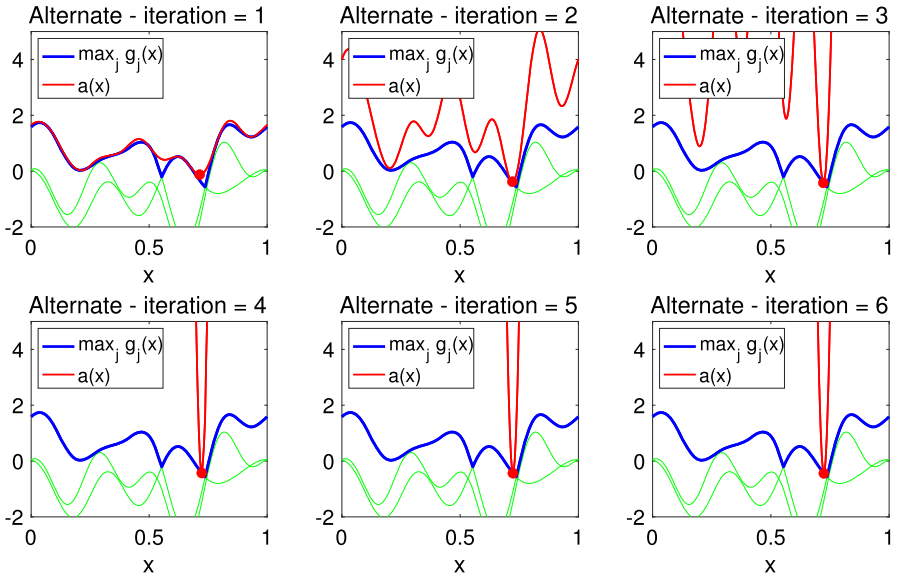

Fig. 5 Two-stage approach for trigonometric polynomials in one dimension, with alternating optimization and $r = 2$, with 6 iterations

the matrix Σ corresponds to a Dirac measure, and the upper-bounding polynomial is so that its value at this point is minimized. This is shown empirically in Fig. 5: even in the good attraction basin, the alternating optimization does not lead to the global optimum.

References

1. Lasserre, J.-B.: Min–max and robust polynomial optimization. J. Glob. Optim. **51**(1), 1–10 (2011)
2. Ben-Tal, A., El Ghaoui, L., Nemirovski, A.: Robust Optimization, vol. 28. Princeton University Press, Princeton (2009)
3. Nie, J., Yang, Z., Zhou, G.: The saddle point problem of polynomials. Found. Comput. Math. 1–37 (2021)
4. Lasserre, J.-B.: Global optimization with polynomials and the problem of moments. SIAM J. Optim. **11**(3), 796–817 (2001)
5. Parrilo, P.A.: Semidefinite programming relaxations for semialgebraic problems. Math. Program. **96**(2), 293–320 (2003)
6. Sion, M.: On general minimax theorems. Pac. J. Math. **8**(1), 171–176 (1958)
7. Jahn, J.: Introduction to the Theory of Nonlinear Optimization. Springer, New York (2020)
8. Laraki, R., Lasserre, J.-B.: Semidefinite programming for min-max problems and games. Math. Program. **131**, 305–332 (2012)
9. Lasserre, J.-B.: Moments, Positive Polynomials and Their Applications, vol. 1. World Scientific, Singapore (2010)
10. Henrion, D., Korda, M., Lasserre, J.-B.: The Moment-SOS Hierarchy: Lectures in Probability, Statistics, Computational Geometry, Control and Nonlinear PDEs. World Scientific, Singapore (2020)
11. Bach, F., Rudi, A.: Exponential convergence of sum-of-squares hierarchies for trigonometric polynomials. Technical report, arXiv (2022)
12. Efthimiou, C.S., Frye, C.: Spherical Harmonics in p Dimensions. World Scientific, Singapore (2014)

13. Schmüdgen, K.: The Moment Problem. Springer, Berlin (2017)
14. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
15. Lofberg, J., Parrilo, P.A.: From coefficients to samples: a new approach to SOS optimization. In: Conference on Decision and Control, vol. 3, pp. 3154–3159 (2004)
16. Rudi, A., Marteau-Ferey, U., Bach, F.: Finding global minima via kernel approximations. Technical Report 2012.11978, arXiv (2020)
17. Morokoff, W.J., Caflisch, R.E.: Quasi-random sequences and their discrepancies. SIAM J. Sci. Comput. **15**(6), 1251–1279 (1994)
18. Helmberg, C., Rendl, F., Vanderbei, R.J., Wolkowicz, H.: An interior-point method for semidefinite programming. SIAM J. Optim. **6**(2), 342–361 (1996)
19. Fazel, M., Hindi, H., Boyd, S.P.: A rank minimization heuristic with application to minimum order system approximation. In: Proceedings of the American Control Conference, vol. 6, pp. 4734–4739 (2001)
20. Henrion, D., Lasserre, J.-B.: Detecting global optimality and extracting solutions in Gloptipoly. Positive Polyn. Control **312**, 293–310 (2005)
21. Fang, K., Fawzi, H.: The sum-of-squares hierarchy on the sphere and applications in quantum information theory. Math. Program. **190**(1), 331–360 (2021)
22. Laurent, M., Slot, L.: An effective version of Schmüdgen's Positivstellensatz for the hypercube. Optim. Lett. 1–16 (2022)
23. Scherer, C.W., Hol, C.W.J.: Matrix sum-of-squares relaxations for robust semi-definite programs. Math. Program. **107**(1–2), 189–211 (2006)
24. Muzellec, B., Bach, F., Rudi, A.: Learning PSD-valued functions using kernel sums-of-squares. Technical Report 2111.11306, arXiv (2021)
25. Golub, G.H., Loan, C.F.V.: Matrix Computations. Johns Hopkins University Press, Baltimore (1996)
26. Ganzburg, M.I.: Multidimensional Jackson theorems. Sib. Math. J. **22**(2), 223–231 (1981)
27. Yu, Y.-L.: The strong convexity of von Neumann's entropy. Unpublished note (2013). http://www.cs.cmu.edu/~yaoliang/mynotes/sc.pdf
28. Lemaréchal, C., Sagastizábal, C.: Practical aspects of the Moreau–Yosida regularization: Theoretical preliminaries. SIAM J. Optim. **7**(2), 367–385 (1997)
29. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley, New Jersey (1999)
30. De Klerk, E., Laurent, M.: On the Lasserre hierarchy of semidefinite programming relaxations of convex polynomial optimization problems. SIAM J. Optim. **21**(3), 824–832 (2011)
31. Putinar, M.: Positive polynomials on compact semi-algebraic sets. Indiana Univ. Math. J. **42**(3), 969–984 (1993)