
S-SOS: Stochastic Sum-Of-Squares for Parametric Polynomial Optimization

Richard L. Zhu

Department of Computational and Applied Mathematics
University of Chicago
Chicago, IL 60637
richardzhu@uchicago.edu

Mathias Oster*

Institute of Geometry and Practical Mathematics
RWTH Aachen
Aachen, Germany
oster@igpm.rwth-aachen.de

Yuehaw Khoo

Department of Statistics
University of Chicago
Chicago, IL 60637
yuehaw.khoo@uchicago.edu

Abstract

Global polynomial optimization is an important tool across applied mathematics, with many applications in operations research, engineering, and the physical sciences. In various settings, the polynomials depend on external parameters that may be random. We discuss a stochastic sum-of-squares (S-SOS) algorithm based on the sum-of-squares hierarchy that constructs a series of semidefinite programs to jointly find strict lower bounds on the global minimum and extract candidates for parameterized global minimizers. We prove quantitative convergence of the hierarchy as the degree increases and use it to solve unconstrained and constrained polynomial optimization problems parameterized by random variables. By employing n -body priors from condensed matter physics to induce sparsity, we can use S-SOS to produce solutions and uncertainty intervals for sensor network localization problems containing up to 40 variables and semidefinite matrix sizes surpassing 800×800 .

1 Introduction

Many effective nonlinear and nonconvex optimization techniques use local information to identify local minima. But it is often the case that we want to find global optima. Sum-of-squares (SOS) optimization is a powerful and general technique in this setting.

The core idea is as follows: suppose we are given polynomials g_1, \dots, g_m, f where each function is on $\mathbb{R}^n \rightarrow \mathbb{R}$ and we seek to determine the minimum value of f on the closed set $\mathcal{S}: \mathcal{S} = \{x \in \mathbb{R}^n \mid g_i(x) \geq 0 \forall i = 1, \dots, m\}$. Our optimization problem is then to find $\inf_{x \in \mathbb{R}^n} \{f(x) \mid x \in \mathcal{S}\}$.

*Corresponding Author

An equivalent formulation is to find the largest constant $c \in \mathbb{R}$ (i.e. the tightest lower bound) that can be subtracted from f such that $f - c \geq 0$ over the set \mathcal{S} . This reduction converts a polynomial optimization problem over a semialgebraic set to the problem of checking polynomial non-negativity. This problem is NP-hard in general [1], therefore one instead resorts to checking if $f - c$ is a sum-of-squares (SOS) function, e.g. in the unconstrained setting where $\mathcal{S} = \mathbb{R}^n$ one seeks to find some polynomials $h_k : \mathbb{R}^n \rightarrow \mathbb{R}$ such that $f - c = \sum_k h_k^2$. If such a decomposition can be found, then we have an easily-checkable certification that $f - c \geq 0$, as all sum-of-squares are non-negative. Since the converse is not true (not all non-negative functions are sum-of-squares), this is a relaxation of the original non-negativity problem.

Notably, if we restrict the h_k to have maximum degree d , the search for a degree- $2d$ SOS decomposition of a function can be automated as a semidefinite program (SDP) [2, 3, 4]. Solving this SDP for varying degrees d generates the well-known Lasserre (SOS) hierarchy. A given degree d corresponds to a particular level of the hierarchy. Solving this SDP produces a lower bound c_d which has been proven to converge to the true global minimum $c^* = \inf_x f(x)$ as d increases, with finite convergence ($c_d = c^*$ at finite d) for functions with second-order local optimality conditions [5, 6] and asymptotic convergence with milder assumptions thanks to representation theorems for positive polynomials from real algebraic geometry [7, 8, 9]. Further work has elucidated both theoretical implications [7, 3, 10, 11, 12, 13] and useful applications of SOS to disparate fields [14, 15, 16, 5, 6, 17, 18] (see further discussion in Appendix A.2).

Motivated by the sum-of-squares certification for a lower bound c on a function $f(x)$, we generalize to the case where the function to be minimized has additional parameters, i.e. $f(x, \omega)$ where x are variables and ω are parameters drawn from some probability distribution $\omega \sim \nu(\omega)$. We seek a function $c(\omega)$ that is the tightest lower bound to $f(x, \omega)$ everywhere: $f(x, \omega) \geq c(\omega)$ with $c(\omega) \rightarrow \inf_x f(x, \omega)$. This setting was originally presented in [19] as a ‘‘Joint and Marginal’’ approach to parametric polynomial optimization. With the view that $\omega \sim \nu(\omega)$ and seeking to parameterize the minimizers $x^*(\omega) = \operatorname{argmin}_x f(x, \omega)$, we are reminded of some of the prior work in polynomial chaos, where a system of stochastic variables is expanded into a deterministic function of those stochastic variables [20, 21].

Contributions and outline. Our primary contributions are a quantitative convergence proof for the Stochastic Sum-of-Squares (S-SOS) hierarchy of semidefinite programs (SDPs), a formulation of a new hierarchy (the cluster basis hierarchy) that uses the structure of a problem to sparsify the SDP, and numerical results on its application to the sensor network localization problem.

In Section 2, we review the S-SOS hierarchy of SDPs [19] and its primal and dual formulations (Section 2.1). We then detail how different hierarchies can be constructed (Section 2.2.1). Finally, in Section 2.3 (complete proof in Appendix A.5.2) we specialize to compact $X \times \Omega$ and outline the proof for a theorem on quantitative convergence (the gap between the optimal values of the degree- $2s$ S-SOS SDP and the ‘‘tightest lower-bounding’’ optimization problem goes $\rightarrow 0$ as $s \rightarrow \infty$) of the S-SOS hierarchy for trigonometric polynomials on $[0, 1]^n \times [0, 1]^d$ following the kernel formalism of [22, 6, 23].

In Section 3 we review the hierarchy’s applications in parametric polynomial minimization and uncertainty quantification, focusing on several variants of sensor network localization on $X \times \Omega = [-1, 1]^n \times [-1, 1]^d$. We present numerical results for the accuracy of the extracted solutions that result from S-SOS, comparing to other approaches to parametric polynomial optimization, including a simple Monte Carlo-based method.

2 Stochastic Sum-of-squares (S-SOS)

2.0.1 Notation

Let $\mathcal{P}(S)$ be the space of polynomials on S , where $S \in \{X, \Omega\}$. $X \subseteq \mathbb{R}^n$ and $\Omega \subseteq \mathbb{R}^d$, respectively, where X and Ω are (not-necessarily compact) subsets of their respective ambient spaces \mathbb{R}^n and \mathbb{R}^d . A polynomial in $\mathcal{P}(X)$ can be written as $p(x) = \sum_{\alpha \in \mathbb{Z}_{\geq 0}^n} c_\alpha x^\alpha \in \mathcal{P}(X)$ (substituting $n \rightarrow d, x \rightarrow \omega, X \rightarrow \Omega$ for a polynomial in $\mathcal{P}(\Omega)$). Let $x := (x_1, \dots, x_n), \omega := (\omega_1, \dots, \omega_d)$, α be a multi-index (size given by context), and c_α be the polynomial coefficients. Let $\mathcal{P}^s(S)$ for some $s \in \mathbb{Z}_{\geq 0}, S \in \{X, \Omega\}$ denote the subspace of $\mathcal{P}(S)$ consisting of polynomials of degree $\leq s$, i.e. polynomials where the multi-indices of the monomial terms satisfy $\|\alpha\|_1 \leq s$. $\mathcal{P}_{\text{SOS}}(X \times \Omega)$ refers

to the space of polynomials on $X \times \Omega$ that can be expressible as a sum-of-squares in x and ω jointly, and $\mathcal{P}_{\text{SOS}}^s(X \times \Omega)$ be the same space restricted to polynomials of degree $\leq s$. Additionally, $W \succcurlyeq 0$ for a matrix W denotes that W is symmetric positive semidefinite (PSD). Finally, $\mathbb{P}(\Omega)$ denotes the set of Lebesgue probability measures on Ω . For more details, see Appendix A.1.

2.1 Formulation of S-SOS hierarchy

We present two formulations of the S-SOS hierarchy that are dual to each other in the sense of Fenchel duality [24, 25]. The primal problem seeks to find the tightest lower-bounding function and the dual problem seeks to find a minimizing probability distribution. Note that the “tightest lower bound” approach is dual to the “minimizing distribution” approach, otherwise known as a “joint and marginal” moment-based approach originally detailed in [19].

2.1.1 Primal S-SOS: The tightest lower-bounding function

Consider a polynomial $f(x, \omega) : \mathbb{R}^{n+d} \rightarrow \mathbb{R}$ with $x \in X \subseteq \mathbb{R}^n, \omega \in \Omega \subseteq \mathbb{R}^d$ equipped with a probability measure $\nu(\omega)$. We interpret x as our optimization variables and ω as noise parameters, and seek a lower-bounding function $c^*(\omega)$ such that $f(x, \omega) \geq c^*(\omega)$ for all x, ω . In particular, we want the tightest lower bound $c^*(\omega) = \inf_{x \in X} f(x, \omega)$. Note that even when $f(x, \omega)$ is polynomial, the tightest lower bound $c^*(\omega)$ can be non-polynomial. A simple example is the function $f(x, \omega) = (x - \omega)^2 + (\omega x)^2$, which has $c^*(\omega) = \inf_x f(x, \omega) = \omega^4 / (1 + \omega^2)$ (Appendix A.6.1).

For us to select the “best” lower-bounding function, we want to maximize the expectation of the lower-bounding function $c(\omega)$ under $\omega \sim \nu(\omega)$ while requiring $f(x, \omega) - c(\omega) \geq 0$, giving us the following optimization problem over L^1 -integrable lower-bounding functions:

$$\begin{aligned} p^* &= \sup_{c \in L^1(\Omega)} \int c(\omega) d\nu(\omega) \\ \text{s.t.} \quad & f(x, \omega) - c(\omega) \geq 0 \end{aligned} \quad (1)$$

Even if we restricted $c(\omega)$ to be polynomial so that the residual $f(x, \omega) - c(\omega)$ is also polynomial, we would still have a challenging nonconvex optimization problem over non-negative polynomials. In SOS optimization, we take a relaxation and require the residual to be SOS: $f(x, \omega) - c(\omega) \in \mathcal{P}_{\text{SOS}}(X \times \Omega)$. Doing the SOS relaxation of the non-negative Equation (1) and restricting $c(\omega)$, i.e. $f(x, \omega) - c(\omega)$ to polynomials of degree $\leq 2s$ gives us Equation (2), which we call the primal S-SOS degree- $2s$ SDP:

$$\begin{aligned} p_{2s}^* &= \sup_{c \in \mathcal{P}^{2s}(\Omega), W \succcurlyeq 0} \int c(\omega) d\nu(\omega) \\ \text{s.t.} \quad & f(x, \omega) - c(\omega) = m_s(x, \omega)^T W m_s(x, \omega) \end{aligned} \quad (2)$$

where $m_s(x, \omega)$ is a basis function $X \times \Omega \rightarrow \mathbb{R}^{a(n, d, s)}$ containing monomial terms of degree $\leq s$ written as a column vector, and $W \in \mathbb{R}^{a(n, d, s) \times a(n, d, s)}$ a symmetric PSD matrix. Here, $a(n, d, s)$ represents the dimension of the basis function, which depends on the degree s and on the dimensions n, d . For this formulation to find the best degree- $2s$ approximation to the lower-bounding function, we require $g(x, \omega) = m_s(x, \omega)^T W m_s(x, \omega)$ to span $\mathcal{P}^{2s}(X \times \Omega)$. Selecting all combinations of standard monomial terms of degree $\leq s$ suffices and results in a basis function with size $a(n, d, s) = \binom{n+d+s}{s}$.

2.1.2 Dual S-SOS: A minimizing distribution

The formal dual to Equation (1) (proof of duality in Appendix A.5.1) seeks to find a “minimizing distribution” $\mu(x, \omega)$, i.e. a probability distribution that places weight on the minimizers of $f(x, \omega)$ subject to the constraint that the marginal $\mu_X(\omega)$ matches $\nu(\omega)$:

$$\begin{aligned} d^* &= \inf_{\mu \in \mathbb{P}(X \times \Omega)} \int f(x, \omega) d\mu(x, \omega) \\ \text{s.t.} \quad & \int_X d\mu(x, \omega) = \mu_X(\omega) = \nu(\omega) \end{aligned} \quad (3)$$

where we have written $\mathbb{P}(X \times \Omega)$ as the space of joint probability distributions on $X \times \Omega$ and $\mu_X(\omega)$ is the marginal of $\mu(x, \omega)$ with respect to x , obtained via disintegration.

For the primal, we considered polynomials of degree $\leq 2s$. We do the same here. The formal dual becomes a tractable SDP, where the objective turns into moment-minimization and the constraints become moment-matching. Following [3, 15], let $M \in \mathbb{R}^{a(n,d,s) \times a(n,d,s)}$ be the symmetric PSD moment matrix with entries defined as $M_{i,j} = \int_{X \times \Omega} m_s^{(i)}(x, \omega) m_s^{(j)}(x, \omega) d\mu(x, \omega)$ where $m_s^{(i)}(x, \omega)$ is the i -th element of the basis function m_s . Let $y \in \mathbb{R}^{b(n,d,s)}$ be the moment vector of independent moments that completely specifies M , e.g. in the case that we use all standard monomials of degree $\leq s$ and have $a(n, d, s) = \binom{n+d+s}{s}$, then $b(n, d, s) = \binom{n+d+2s}{2s}$. We write $M(y)$ as the moment matrix that is formed from these independent moments. We have $y_{\alpha(i,j)} = \int_{X \times \Omega} m_s^{(i)}(x, \omega) m_s^{(j)}(x, \omega) d\mu(x, \omega)$ where the multi-index $\alpha(i, j) \in \mathbb{Z}_{\geq 0}^{n+d}$ corresponds to the sum of the multi-indices corresponding to the i -th entry and the j -th entry of $m_s(x, \omega)$.

We write $f(x, \omega)$ in terms of the monomials $f(x, \omega) = \sum_{\|\alpha\|_1 \leq 2s} f_{\alpha} [x, \omega]^{\alpha}$, where $[x, \omega]$ is the concatenation of the $n + d$ variables from x, ω and $\alpha \in \mathbb{Z}_{\geq 0}^{n+d}$ is a multi-index. Note that every monomial $[x, \omega]^{\alpha}$ has a corresponding moment y_{α} : $\int [x, \omega]^{\alpha} d\mu(x, \omega) = y_{\alpha}$. We then observe that the integral in the objective reduces to a dot product between the coefficients of f and the moment vector:

$$\int f(x, \omega) d\mu(x, \omega) = \int \sum_{\alpha} f_{\alpha} [x, \omega]^{\alpha} d\mu(x, \omega) = \sum_{\alpha} f_{\alpha} y_{\alpha}$$

After converting the distribution-matching constraint $\mu_X(\omega) = \nu(\omega)$ in (3) into equality constraints on the moments of ω up to degree $2s$, we obtain the following dual S-SOS degree- $2s$ SDP:

$$\begin{aligned} d_{2s}^* &= \inf_{y \in \mathbb{R}^{b(n,d,s)}} \sum_{\|\alpha\|_1 \leq 2s} f_{\alpha} y_{\alpha} \\ \text{s.t. } & M(y) \succeq 0 \\ & y_{\alpha} = m_{\alpha} \quad \forall (\alpha, m_{\alpha}) \in \mathcal{M}_{\nu} \end{aligned} \quad (4)$$

We write \mathcal{M}_{ν} as the set of (α, m_{α}) representing the moment-matching constraints on ω^{α} up to degree- $2s$, i.e. we want to set $\int_{X \times \Omega} \omega^{\alpha} d\mu(x, \omega) = \int_{\Omega} \omega^{\alpha} d\nu(\omega) = m_{\alpha}$ for all multi-indices $\alpha \in \mathbb{Z}_{\geq 0}^d$ with $\|\alpha\|_1 \leq 2s$. There are $\binom{d+2s}{2s}$ multi-indices $\alpha \in \mathbb{Z}_{\geq 0}^{n+d}$, $\|\alpha\|_1 \leq 2s$ where only the d entries associated with ω are non-zero, and therefore the number of moment-matching constraints is $|\mathcal{M}_{\nu}| = \binom{d+2s}{2s}$. Note that the moment matrix $M(y) \in \mathbb{R}^{a(n,d,s) \times a(n,d,s)}$ is a symmetric PSD matrix and is the dual variable to the primal W . Observe also that we require the moments of $\nu(\omega)$ of degree up to $2s$ to be bounded. (4) is often a more convenient form than (2), especially when working with additional equality or inequality constraints, as we will see in Section 3. For concrete examples of the primal and dual SDPs with explicit constraints, see Appendix A.3.

2.2 Variations

In this section, we detail two ways of building a hierarchy, one based on the maximum degree of monomial terms in the basis function (Lasserre) and a novel one based on the maximum number of interactions occurring in the terms of the basis function (cluster basis). To define any SOS hierarchy, we first select a monomial basis. Some examples include the standard monomial basis x_1, \dots, x_n , trigonometric/Fourier 1-periodic monomial basis $\sin x_1, \cos x_1, \dots, \sin x_n, \cos x_n$, or others. Using this basis, we write down a basis function $m(x)$ which comprises some combinations of monomials. Squared linear combinations of the basis functions then span a SOS space of functions: $\mathcal{H} : \{(\sum_i h_i m_i(x))^2\}$.

2.2.1 Standard Lasserre hierarchy

In the Lasserre hierarchy, the basis function $m_s(x)$ is composed of all combinations of monomials up to degree $s \in \mathbb{Z}_{>0}$ and a given level of the hierarchy is set by the maximum degree s . The basis function consists of terms x^{α} with α a multi-index and $\|\alpha\|_1 \leq s$. The degree- $2s$ SOS function space parameterized by this basis function is that spanned by $m_s(x)^T W m_s(x)$ for PSD W , i.e. the

functions that can result from squaring any linear combination of degree- s polynomials that can be generated from our basis $m_s(x)$. As we increase the degree s , our basis function gets larger and our S-SOS SDP objective values converge to the optimal value of the “tightest lower-bounding” problem Equation (1) [19].

2.2.2 Cluster basis hierarchy

In this section, we propose a cluster basis hierarchy, wherein we utilize possible spatial organization of the problem to sparsify the problem and reduce the size of the SDP that must be solved [26, 27]. The cluster basis is a physically motivated prior often used in statistical and condensed matter physics, where we assume that our degrees of freedom can be arrayed in space, with locally close variables interacting strongly (kept in the model) and globally separated variables interacting weakly (ignored). Moreover, one may also keep only the terms with interactions between a small number of degrees of freedom, such as considering only pairwise or triplet interactions between particles.

In the cluster basis hierarchy, a given level of the hierarchy is specified by a 2-tuple (b, t) , the desired body order b and the maximum degree of a single variable t . Body order denotes the maximum number of interacting variables in a given monomial term, e.g. $x_i^a x_j^b x_k^c$ would have body order 3 and total degree $a + b + c$. The basis function $m_{b,t}$ consists of terms x^α with α a multi-index, $\|\alpha\|_0 \leq b$ (at most b interacting variables can occur in a single term), and $\|\alpha\|_\infty \leq t$ (each variable can have up to degree t). The maximum degree of the basis function $m_{b,t}$ is then $s = bt$. If we are to compare $m_{b,t}$ from the cluster basis hierarchy with m_s from the Lasserre hierarchy, we find that even when $bt = s$ we still have strictly fewer terms, e.g. in the case where $b = 2, t = 2, s = 4$ we have m_s containing terms of the form x_i^4 but $m_{b,t}$ only has degree-4 terms of the $x_i^2 x_j^2$.

To expand on this, consider that in the standard Lasserre hierarchy we have $m_s(x)$ containing all monomials of degree $\leq s$ in n variables, or $\binom{n+s}{s}$ terms in total. In the proposed cluster basis hierarchy, $m_{b,t}(x)$ has

$$\sum_{k=0}^b \binom{n}{k} t^k$$

terms. This expression results from the need to sum over body orders k , considering that there are $\binom{n}{k}$ ways to choose k variables and that each selected variable has k possible degrees so there are t^k ways to assign degrees $\leq t$ to these k variables. For fixed $b, t, s = bt$, and $n \gg b, t$ we note that $m_s(x)$ has $O(n^{bt})$ terms while $m_{b,t}(x)$ has $O(n^b)$ terms so the size reduction factor in using the cluster basis asymptotically goes like $n^{b(t-1)}$. As the number of variables n in the problem grow, we have asymptotic dominance in using the cluster basis to reduce the size of the SDP that must be solved. As $bt \rightarrow \infty$ we might expect asymptotic convergence of the SDP hierarchy just like the standard Lasserre hierarchy, however, a full proof of convergence is out of scope for this paper. For further details, see discussion in Appendix A.7.4.

2.3 Convergence of S-SOS

As we increase the degree s (either s in the Lasserre hierarchy or b, t in the cluster basis hierarchy) we would expect the SDP objective values p_{2s}^* (Equation (2)) to converge to the optimal value p^* and the lower bounding function $c_{2s}^*(\omega)$ to converge to the tightest lower bound $c^*(\omega) = \inf_x f(x, \omega)$. In this paper we refer to $p_{2s}^* \rightarrow p^*$ and $d_{2s}^* \rightarrow d^*$ interchangeably as strong duality occurs in practice despite being difficult to formally verify (Appendix A.4). This convergence is a common feature of SOS hierarchies.

In this section we show that using polynomial $c_{2s}^*(\omega)$ to approximate $c^*(\omega)$ still allows for asymptotic convergence in L^1 as $s \rightarrow \infty$.

2.3.1 Overview of result

We specialize to the particular case of 1-periodic trigonometric polynomials $f(x, \omega), c(\omega)$ on compact $X = [0, 1]^n$ and compact $\Omega \subset \mathbb{R}^d$ and prove asymptotic convergence of the degree- $2s$ S-SOS hierarchy as $s \rightarrow \infty$. Though seemingly restrictive, this choice allows us to leverage Fourier convergence results on a compact domain. For generic $f(x, \omega)$ where we are only interested in its behavior in some compact set (nearly all practical problems involve a restriction of domain), we may

rescale the compact set to this 1-periodic compact domain and apply the results found here. To use other families of polynomials, we note that a substitution argument justifies the focus on trigonometric polynomials as any convergence result achieved for the trigonometric polynomial hierarchy lead directly to a matching result on regular polynomials (2.2 in [28]). Previously, rates of $1/s^2$ were attained for the standard Lasserre hierarchy (i.e. without external parameters ω) on compact domains with mild assumptions [29, 28, 9]. In our work, a slower convergence rate of $\ln s/s$ is achieved due to the introduction of external parameters, requiring us to approximate the lower-bounding function $c^*(\omega)$ and thereby limiting the obtained convergence rate.

2.3.2 $\ln s/s$ convergence using a polynomial approximation to $c^*(\omega)$

We would like to bound the gap between the optimal lower bound $c^*(\omega) = \inf_{x \in X} f(x, \omega)$ and the lower bound $c_{2s}^*(\omega)$ resulting from solving the degree- $2s$ primal S-SOS SDP, i.e.

$$0 \leq c^*(\omega) - c_{2s}^*(\omega) \leq \varepsilon(f, s) \quad \forall \omega \in \Omega. \quad (5)$$

To that end, we need to understand the regularity of c^* . Without further assumptions, we may assume c^* to be Lipschitz continuous, per Proposition 2.1.

With Equation (5) we may then integrate

$$0 \leq \int_{\Omega} \inf_x f(x, \omega) - c_{2s}^*(\omega) d\nu(\omega) \leq |\Omega| \varepsilon(f, s)$$

where we control ε in terms of the degree s . If we can drive $\varepsilon \rightarrow 0$ as $s \rightarrow \infty$ then we are done.

Proposition 2.1 (Theorem 2.1 in [30]). *Let $g : X \times Y \rightarrow \mathbb{R}$ be polynomial. Then $y \mapsto \inf_{x \in X} g(x, y)$ is Lipschitz continuous.*

Theorem 2.1 (Asymptotic convergence of S-SOS). *Let $f : [0, 1]^n \times \Omega \rightarrow \mathbb{R}$ be a trigonometric polynomial of degree $2r$, $c^*(\omega) = \inf_x f(x, \omega)$ the optimal lower bound as a function of ω , and ν any probability measure on compact $\Omega \subset \mathbb{R}^d$. Let s refer to the degree of the basis in both x, ω terms and the degree of the lower-bounding polynomial $c(\omega)$, i.e. $m_s([x, \omega]) : \mathbb{R}^{n+d} \rightarrow \mathbb{R}^{a(n,d,s)}$ is the full basis function of terms $[x, \omega]^\alpha$ with $\|\alpha\|_1 \leq s$ and $c(\omega)$ only has terms ω^α with $\|\alpha\|_1 \leq s$.*

Let p_{2s}^ be the solution to the following S-SOS SDP (c.f. Equation (2)) with $m_s(x, \omega)$ a spanning basis of trigonometric monomials with degree $\leq s$:*

$$p_{2s}^* = \sup_{c \in \mathcal{P}^{2s}(\Omega), W \succcurlyeq 0} \int c(\omega) d\nu(\omega)$$

$$\text{s.t. } f(x, \omega) - c(\omega) = m_s(x, \omega)^T W m_s(x, \omega)$$

Then there is a constant $C > 0$ depending only on $\|f - \bar{f}\|_F, \|c^ - \bar{c}^*\|, r, \Omega, n, d$ such that the following holds:*

$$\int_{\Omega} [c^*(\omega) - c_{2s}^*(\omega)] d\nu(\omega) \leq C \frac{\ln s}{s}$$

where \bar{f} denotes the average value of the function f over $[0, 1]^n$, i.e. $\bar{f} = \int_{[0, 1]^n} f(x) dx$ and $\|f(x)\|_F = \sum_{\hat{x}} |\hat{f}(\hat{x})|$ denotes the norm of the Fourier coefficients. Thus we have asymptotic convergence of the S-SOS SDP hierarchy to the optimal value p^ of Equation (1) as we send $s \rightarrow \infty$.*

Proof. The following is an outline of the proof. For complete details, including the full theorem and proof, please see Appendix A.5.2.

The key argument used follows that of [22, 28, 23, 31, 32]. Take the lower-bounding minimizer of Equation (1) $c^*(\omega)$ (not necessarily polynomial) and approximate it with a trigonometric polynomial $c_a^*(\omega)$ of degree s . We then pass the non-negative component $f - c_a^*$ through an integral operator T that is built out of sums of squares of trigonometric polynomials (related to the Christoffel-Darboux and Jackson kernels) to obtain a strictly-positive SOS component:

$$Th(x, \omega) = \int_{X \times \Omega} |q_x(x - \bar{x})|^2 |q_\omega(\omega - \bar{\omega})|^2 h(\bar{x}, \bar{\omega}) d\bar{x} d\bar{\omega}.$$

$q_x(x), q_\omega(\omega)$ can be chosen to be “kernel functions” of bounded degree $\leq s$ so that the output SOS function is of degree $2s$. We can show the approximation error $\|c^* - c_a^*\|$ is small, that the operator T exists and is close to the identity, and the deformation resulting from the SOS projection is small for sufficiently large degree s (degree of the approximating $c_a^*(\omega)$ and in the SOS kernel). We may conclude that the true strictly-positive part may be well-approximated by the SOS hierarchy and find asymptotic convergence along with a convergence rate in the degree s of the hierarchy. \square

3 Numerical experiments

We present two numerical studies of S-SOS demonstrating its use in applications. The first study (Section 3.1) numerically tests how the optimal values of the SDP Equation (2) p_{2s}^* converge to p^* of the original primal Equation (1) as we increase the degree. The second study (Section 3.2) evaluates the performance of S-SOS for solution extraction and uncertainty quantification in various sensor network localization problems.

3.1 Simple quadratic SOS function

As a simple illustration of S-SOS, we test it on the SOS function

$$f(x, \omega) = (x - \omega)^2 + (\omega x)^2 \quad (6)$$

with $x \in \mathbb{R}, \omega \in \mathbb{R}$. The lower bound $c^*(\omega) = \inf_x f(x, \omega)$ can be computed analytically as $c^*(\omega) = \omega^4 / (1 + \omega^2)$. Assuming $\omega \sim \text{Uniform}(-1, 1)$, we get that the objective value for the “tightest lower-bounding” primal problem Equation (1) is $p^* = \int_{-1}^1 \frac{\omega^4}{2(1+\omega^2)} d\omega = \frac{\pi}{4} - \frac{2}{3} \approx 0.1187$. For further details, see Appendix A.6.

We are interested in studying the quantitative convergence of the S-SOS hierarchy numerically. The idea is to solve the primal (dual) degree- $2s$ SDP to find the tightest polynomial lower bound (the minimizing probability distribution) for varying degrees s . As s gets larger, the basis function $m_s(x)$ gets larger and the objective value of the SDP Equation (2) p_{2s}^* should converge to the theoretical optimal value p^* .

In Figure 1 we see very good agreement between p^* and p_{2s}^* with exponential convergence as s increases. This is much faster than the rate we found in Section 2.3.2, but agrees with the exponential convergence results from [6] achieved with local optimality assumptions. Due to the simplicity of (6), it is not surprising that we see much faster convergence. In fact, for most typical functions, we might expect convergence much faster than the worst-case rate. The tapering-off of the convergence rate is likely attributed to the numerical tolerance used in our solver (CVXPY/MOSEK), as we observed that increasing the tolerance shifts the best-achieved gap higher.

3.2 Sensor network localization

Sensor network localization (SNL) is a common testbed for global optimization and SDP solvers due to the high sensitivity and ill-conditioning of the problem. In SNL, one seeks to recover the positions of N sensors $X \in \mathbb{R}^{N \times \ell}$ positioned in \mathbb{R}^ℓ given a set of noisy observations of pairwise distances $d_{ij} = \|x_i - x_j\|$ between the sensors [15, 33]. To have a unique global minimum and remove symmetries, sensor-anchor distance observations are often added, where several sensors are anchored at known locations in the space. This can improve the conditioning of the problem, making it “easier” in some sense.

3.2.1 Definitions

We define a SNL *problem instance* with $X \in [-1, 1]^{N \times \ell}$ as the ground-truth positions for $\mathcal{S} = \{1, 2, \dots, N\}$ sensors, $A \in [-1, 1]^{K \times \ell}$ as the ground-truth positions for $\mathcal{A} = \{1, 2, \dots, K\}$ anchors, $\mathcal{D}_{ss}(r) = \{d_{ij} = \|x_i - x_j\| : i, j \in \mathcal{S} \text{ and } d_{ij} \leq r\}$ as the set of observed sensor-sensor distances and $\mathcal{D}_{sa}(r) = \{d_{ik} = \|x_i - a_k\| : i \in \mathcal{S}, k \in \mathcal{A} \text{ and } d_{ik} \leq r\}$ as the set of observed sensor-anchor distances, both of which depend on some sensing radius r .

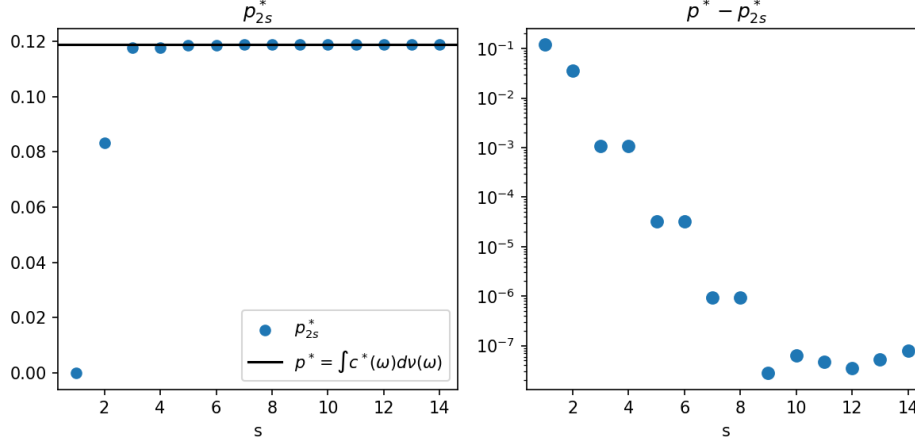


Figure 1: Comparison between the objective value p_{2s}^* from solving the degree- $2s$ S-SOS SDP and the objective value p^* resulting from the best-possible lower bound $c^*(\omega)$ for noise drawn as $\omega \sim \text{Uniform}(-1, 1)$. $p^* = \int c^*(\omega) d\nu(\omega) = \frac{\pi}{4} - \frac{2}{3} \approx 0.1187$ is plotted as the line in black and the p_{2s}^* values are shown as blue dots (left) with the gap between the values $p^* - p_{2s}^*$ (right).

Writing $x_i, a_k \in [-1, 1]^\ell$ as the unknown positions of the i -th sensor and the k -th anchor, we can write the potential function to be minimized as a polynomial:

$$f(x, \omega; X, A, r) = \underbrace{\sum_{d_{ij} \in \mathcal{D}_{ss}(r)} (\|x_i - x_j\|_2^2 - d_{ij}(\omega))^2}_{\text{sensor-sensor interactions}} + \underbrace{\sum_{d_{ik} \in \mathcal{D}_{sa}(r)} (\|x_i - a_k\|_2^2 - d_{ik}(\omega))^2}_{\text{sensor-anchor interactions}} \quad (7)$$

The observed sensor-sensor and sensor-anchor distances $d_{ij}(\omega), d_{ik}(\omega)$ can be perturbed arbitrarily, but in this paper we focus on linear uniform noise, i.e. for a subset of observed distances we have $d_{ij,k}(\omega) = d_{ij,k}^* + \epsilon\omega_k$ with $\omega_k \sim \text{Uniform}(-1, 1)$. Other noise types may be explored, including those including outliers, which may be a better fit for robust methods (Appendix A.7.2).

Equation (7) contains soft penalty terms for sensor-sensor terms and sensor-anchor terms. We can see that this is a degree-4 polynomial in the standard monomial basis elements, and a global minimum of this function is achieved at $f(X, \mathbf{0}^d; X, A, r) = 0$ (where the distances have not been perturbed by any noise). In general for non-zero ω (measuring distances under noise perturbations) we expect the function minimum to be > 0 , as there may not exist a configuration of sensors \hat{X} that is consistent with the observed noisy distances.

We can also support equality constraints in our solution, in particular hard equality constraints on the positions of certain sensors relative to known anchors. This corresponds to removing all sensor-anchor soft penalty terms from the function and instead selecting $N_H < N$ sensors at random to exactly fix in known positions via equality constraints in the SDP. The SDP is still large but the effective number of variable sensors has been reduced to $N' = N - N_H$.

A given SNL *problem type* is specified by a spatial dimension ℓ , N sensors, K anchors, a sensing radius $r \in (0, 2\sqrt{\ell})$, a noise type (linear), and anchor type (soft penalty or hard equality). Once these are specified, we generate a random *problem instance* by sampling $X \sim \text{Uniform}(-1, 1)^n$, $A \sim \text{Uniform}(-1, 1)^d$. The potential $f(x, \omega)$ for a given instance is formed (either with sensor-anchor terms or not, with terms kept based on some sensing radius r , and noise variables appropriately added).

The number of anchors is chosen to be as few as possible so as to still enable exact localization, i.e. $K = \ell + 1$ anchors for a SNL problem in ℓ spatial dimensions. The SDPs are formulated with the help of SymPy [34] and solved using CVXPY [35, 36] and Mosek [37] on a server with two Intel Xeon 6130 Gold processors (32 physical cores total) and 256GB of RAM. For an expanded discussion and further details, see Appendix A.7.

3.2.2 Evaluation metrics

The accuracy of the recovered solution is of primary interest, i.e. our primary evaluation metric should be the distance between our extracted sensor positions x and the ground-truth sensor positions X , i.e. $\text{dist}(x, X)$. Because the S-SOS hierarchy recovers estimates of the sensor positions $\mathbb{E}[x_i]$ along with uncertainty estimates $\text{Var}[x_i]$, we would like to measure the distance between our ground-truth positions X to our estimated distribution $p(x) = \mathcal{N}(\mathbb{E}[x], \text{Var}[x])$. The Mahalanobis distance δ_M (Equation (8)) is a modified distance metric that accounts for the uncertainty [38]. We use this as our primary metric for sensor recovery accuracy.

$$\delta_M(X, \mathcal{N}(\mu, \Sigma)) := \sqrt{(X - \mu)^T \Sigma^{-1} (X - \mu)} \quad (8)$$

As our baseline method, for each problem instance we apply a basic Monte Carlo method detailed in Algorithm 1 (Appendix A.7.3) where we sample $\omega \sim \nu(\omega)$, use a local optimization solver to find $x^*(\omega) = \inf_x f(x, \omega)$, and use this to estimate $\mathbb{E}_{\omega \sim \nu}[x]$, $\text{Var}_{\omega \sim \nu}[x]$. Note that though this non-SOS method achieves some estimate of the dual SDP objective $\int f(x, \omega) d\mu(x, \omega)$, it is not guaranteed to be a lower bound.

3.2.3 Results

Recovery accuracy. In Table 1 we see a comparison of the S-SOS method and the MCPO baseline. Each row corresponds to one SNL problem type, i.e. we fix the physical dimension ℓ , the number of anchors $K = \ell + 1$, and select the sensing radius r and the noise scale ϵ . We then generate $L = 20$ random instances of each problem type, corresponding to a random realization of the ground-truth sensor and anchor configurations $X \in [-1, 1]^{N \times \ell}$, $A \in [-1, 1]^{K \times \ell}$, producing a $f(x, \omega)$ that we then solve the SDP for (in the case of S-SOS) or do pointwise optimizations for (in the case of MCPO). Each method outputs estimates for the sensor positions and uncertainty around it as a $\mathcal{N}(\mathbb{E}[x], \text{Cov}[x])$, which we then compute δ_M for (see Equation (8)), treating each dimension as independent of each other (i.e. X as a flat vector). Each instance solve gives us one observation of δ_M or each method, and we report the median and the $\pm 1\sigma_{34\%}$ values over the $L = 20$ instances we generate.

4 Discussion

In this paper, we discuss the stochastic sum-of-squares (S-SOS) method to solve global polynomial optimization in the presence of noise, prove two asymptotic convergence results for polynomial f and compact Ω , and demonstrate its application to parametric polynomial minimization and uncertainty quantification along with a new cluster basis hierarchy that enables S-SOS to scale to larger problems. In our experiments, we specialized to sensor network localization and low-dimensional uniform random noise with small n, d . However, it is relatively straightforward to extend this method to support other noise types (such as Gaussian random variates without compact support, which we do in Appendix A.6.4) and support higher-dimensional noise with $d \gg 1$.

Scaling this method to larger problems $n \gg 1$ is an open problem for all SOS-type methods. In this paper, we take the approach of sparsification, by making the cluster basis assumption to build up a block-sparse W . We anticipate that methods that leverage sparsity or other structure in f will be promising avenues of research, as well as approximate solving methods that avoid the explicit materialization of the matrices W, M . For example, we assume that the ground-truth polynomial possesses the block-sparse structure because our SDP explicitly requires the polynomial $f(x, \omega)$ to exactly decompose into some lower-bounding $c(\omega)$ and SOS $f_{\text{SOS}}(x, \omega)$. Relaxing this exact-decomposition assumption and generalizing beyond polynomial $f(x, \omega), c(\omega)$ may require novel approaches and would be an exciting area for future work.

Table 1: Comparison of S-SOS and MCPO solution extraction accuracy. We present the Mahalanobis distance δ_M (Equation (8)) of the the true sensor positions X^* to the extracted distribution $\mathcal{N}(\mathbb{E}[x], \text{Var}[x])$ over solutions recovered from S-SOS for varying SNL problem types. ℓ is the spatial dimension, r is the sensing radius used to cutoff terms in the potential $f(x, \omega)$, ϵ is the noise scale, N_H is the number of hard equality constraints used (sensors fixed at known locations), N_C is the number of clusters used (see Appendix A.7.4), and N is the number of sensors used. Each SNL problem instance has $K = \ell + 1$ anchors used in the potential (if $N_H = 0$). The MCPO values are estimated with $T = 50$ Monte Carlo iterates. Each entry is $\hat{\mu} \pm \hat{\sigma}$ where $\hat{\mu}$ is the median and robust standard-deviation ($\sigma_{34\%}$) estimated over 20 runs of the same problem type with varying random initializations of the sensor positions. The entries with the lowest median δ_M are bolded. We also compare the number of elements in the full basis a_f , the cluster basis a_c , and the reduction multiple when using the cluster basis a_f/a_c . When passing to the cluster basis, a_f/a_c is how much the semidefinite matrix shrinks by.

Parameters						Basis comparison			M-distance (δ_M)	
ℓ	r	ϵ	N_H	N_C	N	a_f	a_c	a_f/a_c	S-SOS	MCPO
1	0.5	0.3	0	1	10	78	78	1x	0.94 ± 0.22	2.61 ± 3.86
1	1.0	0.3	0	1	10	78	78	1x	0.29 ± 0.16	1.10 ± 0.58
1	1.5	0.3	0	1	10	78	78	1x	0.11 ± 0.11	0.86 ± 0.52
1	1.5	0.3	2	1	10	78	78	1x	0.24 ± 0.37	1.06 ± 1.28
1	1.5	0.3	4	1	10	78	78	1x	0.10 ± 0.03	0.61 ± 0.41
1	1.5	0.3	6	1	10	78	78	1x	0.06 ± 0.04	0.48 ± 0.32
1	1.5	0.3	8	1	10	78	78	1x	0.04 ± 0.02	0.31 ± 0.17
2	1.5	0.1	0	9	9	406	163	2.5x	2.86 ± 0.94	1562.39 ± 596.29
2	1.5	0.1	0	9	15	820	317	2.6x	3.25 ± 1.19	1848.65 ± 650.45

References

- [1] Michael R. Garey and David S. Johnson. *Computers and intractability: a guide to the theory of NP-completeness*. A series of books in the mathematical sciences. Freeman, New York [u.a], 27. print edition, 2009.
- [2] Yurii Nesterov. Squared Functional Systems and Optimization Problems. In Hans Frenk, Kees Roos, Tamás Terlaky, and Shuzhong Zhang, editors, *High Performance Optimization*, Applied Optimization, pages 405–440. Springer US, Boston, MA, 2000.
- [3] Jean B. Lasserre. Global Optimization with Polynomials and the Problem of Moments. *SIAM Journal on Optimization*, 11(3):796–817, January 2001. Publisher: Society for Industrial and Applied Mathematics.
- [4] Monique Laurent. Sums of Squares, Moment Matrices and Optimization Over Polynomials. In Mihai Putinar and Seth Sullivant, editors, *Emerging Applications of Algebraic Geometry*, The IMA Volumes in Mathematics and its Applications, pages 157–270. Springer, New York, NY, 2009.
- [5] Jiawang Nie. Optimality conditions and finite convergence of Lasserre’s hierarchy. *Mathematical Programming*, 146(1-2):97–121, August 2014.
- [6] Francis Bach and Alessandro Rudi. Exponential convergence of sum-of-squares hierarchies for trigonometric polynomials, April 2023. arXiv:2211.04889 [math].
- [7] Mihai Putinar. Positive Polynomials on Compact Semi-algebraic Sets. *Indiana University Mathematics Journal*, 42(3):969–984, 1993. Publisher: Indiana University Mathematics Department.
- [8] Konrad Schmüdgen. *The Moment Problem*, volume 277 of *Graduate Texts in Mathematics*. Springer International Publishing, Cham, 2017.

- [9] Lucas Slot and Monique Laurent. Sum-of-squares hierarchies for binary polynomial optimization. *Mathematical Programming*, 197(2):621–660, January 2023. Publisher: Springer Verlag.
- [10] Jean Lasserre. The Moment-SOS hierarchy, August 2018. arXiv:1808.03446 [math].
- [11] Jean-Bernard Lasserre. The Moment-SOS hierarchy: Applications and related topics. *To appear in Acta Numerica (2024)*, September 2023.
- [12] Lorenzo Baldi and Bernard Mourrain. On the effective Putinar’s Positivstellensatz and moment approximation. *Mathematical Programming*, 200(1):71–103, June 2023.
- [13] Lorenzo Baldi and Lucas Slot. Degree Bounds for Putinar’s Positivstellensatz on the Hypercube. *SIAM Journal on Applied Algebra and Geometry*, 8(1):1–25, March 2024.
- [14] P.A. Parrilo. *Structured Semidefinite Programs and Semialgebraic Geometry Methods in Robustness and Optimization*. PhD thesis, California Institute of Technology, Pasadena, CA, 2000.
- [15] Jiawang Nie. Sum of squares method for sensor network localization. *Computational Optimization and Applications*, 43(2):151–179, June 2009.
- [16] Etienne de Klerk. The complexity of optimizing over a simplex, hypercube or sphere: a short survey. *Central European Journal of Operations Research*, 16(2):111–125, June 2008.
- [17] Amir Ali Ahmadi and Anirudha Majumdar. DSOS and SDSOS Optimization: More Tractable Alternatives to Sum of Squares and Semidefinite Optimization. *SIAM Journal on Applied Algebra and Geometry*, 3(2):193–230, January 2019.
- [18] Dávid Papp and Sercan Yildiz. Sum-of-Squares Optimization without Semidefinite Programming. *SIAM Journal on Optimization*, 29(1):822–851, January 2019.
- [19] Jean B. Lasserre. A “Joint+Marginal” Approach to Parametric Polynomial Optimization. *SIAM Journal on Optimization*, 20(4):1995–2022, January 2010. Publisher: Society for Industrial and Applied Mathematics.
- [20] Bruno Sudret. Global sensitivity analysis using polynomial chaos expansion. *Reliability Engineering & System Safety*, 93:964–979, July 2008.
- [21] Habib N. Najm. Uncertainty Quantification and Polynomial Chaos Techniques in Computational Fluid Dynamics. *Annual Review of Fluid Mechanics*, 41(1):35–52, 2009. _eprint: <https://doi.org/10.1146/annurev.fluid.010908.165248>.
- [22] Kun Fang and Hamza Fawzi. The sum-of-squares hierarchy on the sphere, and applications in quantum information theory. *Mathematical Programming*, 190(1-2):331–360, November 2021. arXiv:1908.05155 [quant-ph].
- [23] Lucas Slot. Sum-of-Squares Hierarchies for Polynomial Optimization and the Christoffel–Darboux Kernel. *SIAM Journal on Optimization*, 32(4):2612–2635, December 2022.
- [24] Ralph Tyrell Rockafellar. *Convex Analysis*. Princeton Landmarks in Mathematics and Physics. Princeton University Press, Princeton, NJ, 2015.
- [25] Stephen Boyd and Lieven Vandenbergh. *Convex Optimization*. Cambridge University Press, 1 edition, March 2004.
- [26] Lieven Vandenbergh. Chordal Graphs and Sparse Semidefinite Optimization, 2017.
- [27] Yian Chen, Yuehaw Khoo, and Lek-Heng Lim. Convex Relaxation for Fokker-Planck, June 2023. arXiv:2306.03292 [cs, math].
- [28] Francis Bach. Sums-of-squares for dummies: a view from the Fourier domain – Machine Learning Research Blog, November 2022.

- [29] Markus Schweighofer. On the complexity of Schmüdgen’s Positivstellensatz. *Journal of Complexity*, 20(4):529–543, August 2004.
- [30] Frank H. Clarke. Generalized gradients and applications. *Transactions of the American Mathematical Society*, 205:247–262, 1975.
- [31] Corbinian Schlosser, Matteo Tacchi, and Alexey Lazarev. Convergence rates for the moment-SoS hierarchy, February 2024. arXiv:2402.00436 [math].
- [32] Milan Korda, Victor Magron, and Rodolfo Ríos-Zertuche. Convergence rates for sums-of-squares hierarchies with correlative sparsity. *Mathematical Programming*, March 2024.
- [33] Anthony Man-Cho So and Yinyu Ye. Theory of semidefinite programming for Sensor Network Localization. *Mathematical Programming*, 109(2-3):367–384, January 2007.
- [34] Aaron Meurer, Christopher P. Smith, Mateusz Paprocki, Ondřej Čertík, Sergey B. Kirpichev, Matthew Rocklin, AMiT Kumar, Sergiu Ivanov, Jason K. Moore, Sartaj Singh, Thilina Rathnayake, Sean Vig, Brian E. Granger, Richard P. Muller, Francesco Bonazzi, Harsh Gupta, Shivam Vats, Fredrik Johansson, Fabian Pedregosa, Matthew J. Curry, Andy R. Terrel, Štěpán Roučka, Ashutosh Saboo, Isuru Fernando, Sumith Kulal, Robert Cimrman, and Anthony Scopatz. Sympy: symbolic computing in python. *PeerJ Computer Science*, 3:e103, January 2017.
- [35] Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.
- [36] Akshay Agrawal, Robin Verschueren, Steven Diamond, and Stephen Boyd. A rewriting system for convex optimization problems. *Journal of Control and Decision*, 5(1):42–60, 2018.
- [37] MOSEK ApS. *The MOSEK optimization toolbox for Python manual. Version 10.0.*, 2023.
- [38] PC Mahalanobis. On the generalized distance in statistics. In *Proceedings National Institute of Science of India*, volume 49, pages 234–256, 1936. Issue: 2.
- [39] Ryan O’Donnell. SOS is not obviously automatizable, even approximately. *R. O.*, 2016.
- [40] Matthew B. Hastings. Perturbation Theory and the Sum of Squares, June 2022. arXiv:2205.12325 [cond-mat, physics:hep-th, physics:quant-ph].
- [41] M. B. Hastings. Field Theory and The Sum-of-Squares for Quantum Systems, February 2023. arXiv:2302.14006 [quant-ph].
- [42] Samuel B. Hopkins. *Statistical Inference and the Sum of Squares Method*. PhD Thesis, Cornell University, August 2018.
- [43] Samuel B. Hopkins and Jerry Li. Mixture models, robustness, and sum of squares proofs. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1021–1034, Los Angeles CA USA, June 2018. ACM.
- [44] Saeid Sedighi, Kumar Vijay Mishra, M. R. Bhavani Shankar, and Bjorn Ottersten. Localization With One-Bit Passive Radars in Narrowband Internet-of-Things Using Multivariate Polynomial Optimization. *IEEE Transactions on Signal Processing*, 69:2525–2540, January 2021. ADS Bibcode: 2021ITSP...69.2525S.
- [45] Dimitris Bertsimas, Dan Andrei Iancu, and Pablo A. Parrilo. A Hierarchy of Near-Optimal Policies for Multistage Adaptive Optimization. *IEEE Transactions on Automatic Control*, 56(12):2809–2824, December 2011.
- [46] Luigi Ambrosio, Nicola Gigli, and Savare. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Birkhäuser, second edition, 2005.
- [47] Ivar Ekeland and Roger Temam. *Convex analysis and variational problems*, 1976.
- [48] D. Jackson. *The Theory of Approximation*. Colloquium Publications. American Mathematical Society, 1930.

- [49] Nathan Krislock and Henry Wolkowicz. Explicit Sensor Network Localization using Semidefinite Representations and Facial Reductions, January 2010. arXiv:1002.0013 [math].
- [50] Brian D. O. Anderson, Peter N. Belhumeur, Tolga Eren, David K. Goldenberg, A. Stephen Morse, Walter Whiteley, and Y. Richard Yang. Graphical properties of easily localizable sensor networks. *Wireless Networks*, 15(2):177–191, February 2009.
- [51] Brian D. O. Anderson, Iman Shames, Guoqiang Mao, and Barış Fidan. Formal Theory of Noisy Sensor Network Localization. *SIAM Journal on Discrete Mathematics*, 24(2):684–698, January 2010.

A Appendix / supplemental material

A.1 Notation

Let $\mathcal{P}(X)$ and $\mathcal{P}(\Omega)$ denote the spaces of polynomials on $X \subseteq \mathbb{R}^n$ and $\Omega \subseteq \mathbb{R}^d$, respectively, where X and Ω are (not-necessarily compact) subsets of their respective ambient spaces \mathbb{R}^n and \mathbb{R}^d . Specifically, all polynomials of the forms below belong to their respective spaces:

$$p(x) = \sum_{\alpha \in \mathbb{Z}_{\geq 0}} c_{\alpha} x^{\alpha} \in \mathcal{P}(X), \quad p(\omega) = \sum_{\alpha \in \mathbb{Z}_{\geq 0}} c_{\alpha} \omega^{\alpha} \in \mathcal{P}(\Omega)$$

where $x = (x_1, \dots, x_n)$, $\omega = (\omega_1, \dots, \omega_d)$, α is a multi-index for the respective spaces, and c_{α} are the polynomial coefficients.

Let $\mathcal{P}^d(S)$ for some $S \in \{X, \Omega\}$ denote the subspace of $\mathcal{P}(S)$ consisting of polynomials of degree $\leq d$, i.e. polynomials where the multi-indices of the monomial terms satisfy $\|\alpha\|_1 \leq d$. $\mathcal{P}_{\text{SOS}}(X \times \Omega)$ refers to the space of polynomials on $X \times \Omega$ that can be expressible as a sum-of-squares in x and ω jointly. Additionally, $W \succcurlyeq 0$ for a matrix W denotes that W is symmetric positive semidefinite (PSD). Finally, $\mathbb{P}(\Omega)$ denotes the set of Lebesgue probability measures on Ω .

A.2 Related work

A.2.1 Sum-of-squares theory and practice

The theoretical justification underlying the SDP relaxations in global optimization we use here derive from the Positivstellensatz (positivity certificate) of [7], a representation theorem guaranteeing that strictly positive polynomials on certain sets admit sum-of-squares representations. Following this, [3, 10, 11] developed the Moment-SOS hierarchy, describing a hierarchy of primal-dual SDPs (each having fixed degree) of increasing size that provides a monotonic non-decreasing sequence of lower bounds.

There is rich theory underlying the SOS hierarchy combining disparate results from algebraic geometry [14, 10, 11], semidefinite programming [15, 18], and complexity theory [16, 39]. The hierarchy exhibits finite convergence in particular cases where convexity and a strict local minimum are guaranteed [5], otherwise converging asymptotically [6]. In practice, the hierarchy often does even better than these guarantees, converging exactly at c_s^* for some small s .

The SOS hierarchy has found numerous applications in wide-ranging fields, including: reproducing certain results of perturbation theory and providing useful lower-bound certifications in quantum field theory and quantum chemistry [40, 41], providing better provable guarantees in high-dimensional statistical problems [42, 43], useful applications in the theory and practice of sensor network localization [15, 44] and in robust and stochastic optimization [45].

Due to the SDP relaxation, the SOS hierarchy is quite powerful. This flexibility comes at a cost, primarily in the form of computational complexity. The SDP prominently features a PSD matrix $W \in \mathbb{R}^{a(n,d,s) \times a(n,d,s)}$ with $a(n,d,s)$ scaling as $\binom{n+d+s}{s}$ for n dimensions and maximum degree s . Without exploiting the structure of the polynomial, such as locality (coupled terms) or sparsity, solving the SDP using a standard interior point method becomes prohibitively expensive for moderate values of s or n . Work attempting to improve the scalability of the core ideas underlying the SOS hierarchy and the SDP method include [17, 18].

A.2.2 Stochastic sum-of-squares and parametric polynomial optimization

The S-SOS hierarchy we present in this paper as a solution to parametric polynomial optimization was presented originally by [19] as a ‘‘Joint + Marginal’’ approach. That work provides the same hierarchy of semidefinite relaxations where the sequence of optimal solutions converges to the moment vector of a probability measure encoding all information about the globally-optimal solutions $x^*(\omega) = \operatorname{argmin}_x f(x, \omega)$ and provides a proof that the dual problem (our primal) obtains a polynomial approximation to the optimal value function that converges almost-uniformly to $c^*(\omega)$.

A.2.3 Uncertainty quantification and polynomial chaos

Once a physical system or optimization problem is characterized, sensitivity analysis and uncertainty quantification seek to quantify how randomness or uncertainty in the inputs can affect the response. In our work, we have the parametric problem of minimizing a function $f(x, \omega)$ over x where ω parameterizes the function and is drawn from some noise distribution $\nu(\omega)$.

If only function evaluations $f(x, \omega)$ are allowed and no other information is known, Monte Carlo is often applied, where one draws $\omega_k \sim \nu(\omega)$ and solves many realizations of $\inf_x f_k(x) = f(x, \omega_k)$ to approximately solve the following stochastic program:

$$f^* = \inf_x \mathbb{E}_{\omega \sim \nu}[f(x, \omega)]$$

Standard Monte Carlo methods are ill-suited for integrating high-dimensional functions, so this method is computationally challenging in its own right. In addition, we have no guarantees on our result except that as we take the number of Monte Carlo iterates $T \rightarrow \infty$ we converge to some unbiased estimate of $\mathbb{E}_{\omega \sim \nu}[f(x, \omega)]$.

Our approach to quantifying the uncertainty in optimal function value resulting from uncertainty in parameters ω is to find a deterministic lower-bounding $c^*(\omega)$ which guarantees $f(x, \omega) \geq c^*(\omega)$ no matter the realization of noise. This is reminiscent of the polynomial chaos expansion literature, wherein a system of some stochastic variables is expanded into a deterministic function of those stochastic variables, usually in some orthogonal polynomial basis [20, 21].

A.3 An example

Example A.1. Let $f(x, \omega)$ be some polynomial of degree $\leq 2s$ written in the standard monomial basis, i.e.

$$\begin{aligned} f(x, \omega) &= \sum_{\|\alpha\|_1 \leq 2s} f_\alpha x^\alpha \\ &= \sum_{\|\alpha\|_1 \leq 2s} f_{(\alpha_1, \dots, \alpha_{n+d})} \prod_{i=1}^n x_i^{\alpha_i} \prod_{i=1}^d \omega_i^{\alpha_{n+i}} \end{aligned}$$

Let $m_s(x, \omega) \in \mathbb{R}^{a(n, d, s)}$ be the basis vector representing the full set of monomials in x, ω of degree $\leq s$ with $a(n, d, s) = \binom{n+d+s}{s}$.

For all $\alpha \in \mathbb{Z}_{\geq 0}^{n+d}$ with $\|\alpha\|_1 \leq 2s$ and $\alpha_k = 0$ for all $k \in \{1, \dots, n\}$ (i.e. monomial terms containing only $\omega_1, \dots, \omega_d$) we must have:

$$\int_{X \times \Omega} \omega^\alpha d\mu(x, \omega) - \int_{\Omega} \omega^\alpha d\nu(\omega) = 0$$

Explicitly, for μ to be a valid probability distribution we must have:

$$\int_{X \times \Omega} d\mu(x, \omega) - 1 = M_{0,0} - 1 = y_{(1,0,\dots)} - 1 = 0$$

Suppose $\Omega = [-1, 1]$, $\omega \sim \text{Uniform}(-1, 1)$ so that $d = 1$, $\nu(\omega) = 1/2$. We require:

$$\int_{X \times \Omega} \omega^\alpha d\mu(x, \omega) = \int_{[-1,1]} \omega^\alpha d\nu(\omega) = \begin{cases} 1 & \alpha = 0 \\ 0 & \alpha = 1 \\ \frac{1}{3} & \alpha = 2 \\ 0 & \alpha = 3 \\ \frac{1}{5} & \alpha = 4 \end{cases}$$

□

A.4 Strong duality

To guarantee strong duality theoretically, we need a strictly feasible point in the interior (Slater's condition). For us, this is a consequence of Putinar's Positivstellensatz, if $f(x, \omega)$ admits a decomposition as $f(x, \omega) = c(\omega) + g(x, \omega)$ where $g(x, \omega) > 0$ (i.e. is strictly positive), we have strong duality, i.e. $p^* = d^*$ and $p_{2s}^* = d_{2s}^*$ [3, 8]. However, it is difficult to verify the conditions analytically. In practice, strong duality is observed in most cases, so in this paper we refer to solving the primal and dual interchangeably, as $p_{2s}^* = d_{2s}^*$ in all cases we encounter where a SDP solver returns a feasible point.

A.5 Proofs

A.5.1 Primal-dual relationship of S-SOS

Regular SOS Global polynomial optimization can be framed as the following lower-bound maximization problem where we need to check global non-negativity:

$$\begin{aligned} \sup_{c \in \mathbb{R}} \quad & c \\ \text{s.t.} \quad & f(x) - c \geq 0 \quad \forall x \end{aligned} \tag{9}$$

When we take the SOS relaxation of the non-negativity constraint in the primal, we now arrive at the SOS primal problem, where we require $f(x) - c$ to be SOS which guarantees non-negativity but is a stronger condition than necessary:

$$\begin{aligned} \sup_{c \in \mathbb{R}} \quad & c \\ \text{s.t.} \quad & f(x) - c \in \mathcal{P}_{\text{SOS}}(X). \end{aligned} \tag{10}$$

The dual to Equation (9) is the following moment-minimization problem:

$$\begin{aligned} \inf_{\mu \in \mathbb{P}(X)} \quad & \int f(x) d\mu(x) \\ \text{with} \quad & \int d\mu(x) = 1. \end{aligned} \tag{11}$$

Taking some spanning basis $m_s(x) : \mathbb{R}^n \rightarrow \mathbb{R}^{a(n,s)}$ of monomials up to degree s , we have the moment matrix $M \in \mathbb{R}^{a(n,s) \times a(n,s)}$:

$$M_{i,j} = \int m_i(x) m_j(x) d\mu(x) = y_\alpha$$

where we introduce a moment vector y whose elements correspond to the unique moments of the matrix M . Then we may write the degree- $2s$ moment-minimization problem, which is now in a solvable numerical form:

$$\begin{aligned} \inf_y \quad & \sum_{\alpha} f_{\alpha} y_{\alpha} \\ \text{with} \quad & M(y)_{1,1} = 1 \\ & M(y) \succcurlyeq 0 \end{aligned} \tag{12}$$

where we write $M(y)$ as the matrix formed by placing the moments from y into their appropriate places and we set the first element of $m_s(x)$ to be 1, hence $M_{1,1} = \int d\nu(x) = 1$ is simply the normalization constraint. For further reading, see [15, 3].

Stochastic SOS Now let us lift this problem into the stochastic setting with parameters ω sampled from a given distribution ν , i.e. replacing $x \rightarrow (x, \omega)$. We need to make some choice for the objective. The expectation of the lower bound under $\nu(\omega)$ is a reasonable choice, i.e.

$$\int_{\Omega} c(\omega) d\nu(\omega)$$

but we could also make other choices, such as ones that encourage more robust lower bounds. In this paper however, we formulate the primal S-SOS as below (same as Equation (1)):

$$\begin{aligned} p^* &= \sup_{c \in L^1(\Omega)} \int c(\omega) d\nu(\omega) \\ \text{s.t. } & f(x, \omega) - c(\omega) \geq 0 \end{aligned} \quad (13)$$

Note that if the ansatz space for the function $c(\omega)$ is general enough, the maximization of the curve c is equivalent to a pointwise maximization, i.e. we recover the best approximation for almost all ω . Then the dual problem has a very similar form to the non-stochastic case.

Theorem A.2. *The dual to Equation (13) is the following moment minimization where $\mu(x, \omega)$ is a probability measure on $X \times \Omega$:*

$$\begin{aligned} & \inf_{\mu \in \mathbb{P}(X \times \Omega)} \int f(x, \omega) d\mu(x, \omega) \\ \text{with } & \int_{X \times \Omega} \omega^\alpha d\mu(x, \omega) = \int_{\Omega} \omega^\alpha d\nu(\omega) \quad \text{for all } \alpha \in \mathbb{N}^d. \end{aligned}$$

Remark A.3. Notice, that the condition $\int_{X \times \Omega} \omega^\alpha d\mu(x, \omega) = \int_{\Omega} \omega^\alpha d\nu(\omega)$ implies that the first marginal of μ is the noise distribution ν . Let μ_ω denote the disintegration of μ with respect to ν , [46]. Then the moment matching condition is equivalent to $\mu_\omega(X) = 1$ for almost all ω and μ being a Young measure w.r.t. ν . The idea is that $\mu_\omega(x)$ is a minimizing density for every single configuration of ω .

Proof. We use $\mathcal{P}_{\geq 0}(X \times \Omega)$ to denote the space of non-negative polynomials on $X \times \Omega$. Given measure ν on Ω and polynomial function $p : X \times \Omega \rightarrow \mathbb{R}$ consider

$$\begin{aligned} & \sup_{\substack{\gamma \in L^1(\Omega, \nu) \\ q \in \mathcal{P}_{\geq 0}(X \times \Omega)}} \int_{\Omega} \gamma(\omega) d\nu(\omega) \\ \text{s.t. } & p(x, \omega) - \gamma(\omega) = q(x, \omega) \end{aligned}$$

This is equivalent to

$$- \inf_{\substack{\gamma \in L^1(\Omega, \mu) \\ q \in \mathcal{P}_{\geq 0}(X \times \Omega)}} f(\gamma, q) + g(\gamma, q)$$

with

$$f(\gamma, q) = - \int_{\Omega} \gamma(\omega) d\nu(\omega)$$

and

$$g(\gamma, q) = -\chi_{\{f - \gamma - q = 0\}} = \begin{cases} 0 & \text{if } f - \gamma - q = 0 \\ -\infty & \text{else} \end{cases},$$

i.e. g is the characteristic function enforcing non-negativity.

Denote by h^* the Legendre dual, i.e.

$$h^*(y) = \sup_x \langle x, y \rangle - h(x).$$

Then by Rockafellar duality, [47, 24], and noting that signed Borel measures \mathcal{B} are the dual to continuous functions, the dual problem reads

$$\sup_{\Gamma \in L^\infty(\Omega, \mu), \mu \in \mathcal{B}} -f^*(\Gamma, \mu) - g^*(-(\Gamma, \nu))$$

and we would have

$$\sup_{\Gamma \in L^\infty(\Omega, \mu), \mu \in \mathcal{B}} -f^*(\Gamma, \mu) - g^*(-(\Gamma, \mu)) = - \inf_{\substack{\gamma \in L^1(\Omega, \mu) \\ q \in \mathcal{P}_{\geq 0}(X \times \Omega)}} f(\gamma, q) + g(\gamma, q).$$

The Legendre duals of f and g can be explicitly calculated as

$$f^*(\Gamma, \mu) = \begin{cases} 0 & \text{if } \Gamma = -1 \text{ and } \mu \leq 0 \\ \infty & \text{else} \end{cases}$$

and

$$g^*(\Gamma, \mu) = \begin{cases} \int_{\Omega \times X} f(x, \omega) d\mu(\omega, x) & \text{if } f - \gamma \in \mathcal{P}_{\geq 0}(X \times \Omega) \text{ and } \Gamma(\omega) = \mu_\omega(X) \\ \infty & \text{else} \end{cases}$$

since

$$\begin{aligned} f^*(\Gamma, \mu) &= \sup_{\gamma, q} \left(\int_{\Omega} \gamma(\omega) \Gamma(\omega) d\nu(\omega) + \int_{\Omega \times X} q(x, \omega) d\mu(x, \omega) - f(\gamma, q) \right) \\ &= \sup_{\gamma, q} \int_{\Omega} \gamma(\omega) (\Gamma(\omega) + 1) d\nu(\omega) + \int_{\Omega \times X} q(x, \omega) d\mu(x, \omega) \\ &= \begin{cases} 0 & \text{if } \Gamma = -1 \text{ and } \mu \leq 0 \\ \infty & \text{else} \end{cases} \end{aligned}$$

and

$$\begin{aligned} g^*(\Gamma, \mu) &= \sup_{\gamma, q} \int_{\Omega} \gamma(\omega) \Gamma(\omega) d\nu(\omega) + \int_{\Omega \times X} q(x, \omega) d\mu(\omega, x) + \chi_{\{f - \gamma - q = 0\}} \\ &= \begin{cases} \sup_{\gamma} \int_{\Omega} \gamma(\omega) \Gamma(\omega) d\nu(\omega) + \int_{\Omega \times X} (f(x, \omega) - \gamma(\omega)) d\mu(\omega, x) & \text{if } f - \gamma \in \mathcal{P}_{\geq 0}(X \times \Omega) \\ \infty & \text{else} \end{cases} \\ &= \begin{cases} \sup_{\gamma} \int_{\Omega} \gamma(\omega) (\Gamma(\omega) - \mu_\omega(X)) d\nu(\omega) + \int_{\Omega \times X} (f(x, \omega) - \gamma(\omega)) d\mu(\omega, x) & \text{if } f - \gamma \in \mathcal{P}_{\geq 0}(X \times \Omega) \\ \infty & \text{else} \end{cases} \\ &= \begin{cases} \int_{\Omega \times X} (f(x, \omega) - \gamma(\omega)) d\mu(\omega, x) & \text{if } f - \gamma \in \mathcal{P}_{\geq 0}(X \times \Omega) \text{ and } \Gamma(\omega) = \mu_\omega(X) \\ \infty & \text{else} \end{cases} \end{aligned}$$

Altogether, we get

$$-f^*(\Gamma, \mu) - g^*(-\Gamma, -\mu) = \begin{cases} \int_{\Omega \times X} f(x, \omega) d\mu(\omega, x) & \text{if } \mu_\omega(X) = 1 \\ \infty & \text{else.} \end{cases}$$

□

A.5.2 Convergence of S-SOS hierarchy

$c^*(\omega)$ is at best Lipschitz continuous By Proposition 2.1, we argue that $c^*(\omega) = \inf_x f(x, \omega)$ is Lipschitz continuous. One cannot expect much more as the following example shows:

Example A.4. Consider $g : \mathbb{R} \times \mathbb{R}^2 \rightarrow \mathbb{R}$ defined by

$$g(x, p, q) = (x^2 + px + q)^2.$$

Then we have for every $(p, q) \in \mathbb{R}^2$ that

$$\inf_{x \in \mathbb{R}} g(x, p, q) = \begin{cases} 0 & \text{if } \frac{p^2}{4} \geq q \\ \left(\frac{p^2}{4} - q\right)^2 & \text{else.} \end{cases}$$

Therefore, $(p, q) \mapsto \inf_{x \in \mathbb{R}} g(x, p, q)$ is once differentiable but not twice. □

Lemma on approximating polynomials

Lemma A.5. *Let Ω be a compact subset of \mathbb{R}^n and $g : \Omega \rightarrow \mathbb{R}$ be Lipschitz continuous. Then there exists a trigonometric polynomial g_s of degree s and a constant $C > 0$ depending only on Ω and n such that*

$$g \geq g_s$$

and

$$\|g - g_s\|_{L^1(\Omega)} \leq \frac{1 + \ln(s)}{s} C \|g\|_{H^1(\Omega)}.$$

Proof. By Jackson's inequality for Lipschitz functions [48] we have the existence of a trigonometric polynomial g' of degree s with

$$\|g - g'\|_{L^1(\Omega)} \leq \frac{C'}{s} \|g\|_{H^1(\Omega)}$$

as well as

$$\|g - g'\|_{L^\infty(\Omega)} \leq L_g \frac{\ln(s)}{s}.$$

Then we define $g_s = g' - \|g - g'\|_{L^\infty(\Omega)}$ and hence $g \geq g_s$. Furthermore,

$$\|g - g_s\|_{L^1(\Omega)} \leq \frac{(C' + |\Omega| \ln s)}{s} L_g.$$

Writing $C(\Omega) = \max\{C', |\Omega|\}$ we have the desired form where $|\Omega|$ is the volume of Ω . \square

Convergence at $\ln s/s$ rate using integral operator methodology

Theorem A.1 (Asymptotic convergence of S-SOS). *Let $f : [0, 1]^n \times \Omega \rightarrow \mathbb{R}$ be a trigonometric polynomial of degree $2r$, $c^*(\omega) = \inf_x f(x, \omega)$ the optimal lower bound as a function of ω , and ν any probability measure on compact $\Omega \subset \mathbb{R}^d$. Let $s = (s_x, s_\omega, s_c)$, referring separately to the degree of the basis in x terms, the degree of the basis in ω terms, and the degree of the lower-bounding polynomial $c(\omega)$.*

Let $c_{2s}^(\omega)$ be the lower bounding function obtained from the primal S-SOS SDP with $m_s(x, \omega)$ a spanning basis of trigonometric monomials with degree $\leq s_x$ in x terms and of degree $\leq s_\omega$ in ω terms:*

$$p_{2s}^* = \sup_{c \in \mathcal{P}^{2s_c}(\Omega), W \succcurlyeq 0} \int c(\omega) d\nu(\omega)$$

$$\text{s.t. } f(x, \omega) - c(\omega) = m_s(x, \omega)^T W m_s(x, \omega)$$

Then there is a constant $C > 0$ depending only on Ω , d , and n such that for all $s_\omega, s_x \geq \max\{3r, 3s_c\}$ the following holds:

$$\int_{\Omega} [c^*(\omega) - c_{2s}^*(\omega)] d\nu(\omega) \leq |\Omega| \epsilon(f, s)$$

$$\epsilon(f, s) \leq \|f - \bar{f}\|_F \left[1 - \left(1 - \frac{6r^2}{s_\omega^2}\right)^{-d} \left(1 - \frac{6r^2}{s_x^2}\right)^{-n} \right]$$

$$+ \|c^* - \bar{c}^*\|_F \left[1 - \left(1 - \frac{6r^2}{s_\omega^2}\right)^{-d} \right] + C \frac{(1 + \ln(2s_c))}{2s_c}.$$

where \bar{f} denotes the average value of the function f over $[0, 1]^n$, i.e. $\bar{f} = \int_{[0, 1]^n} f(x) dx$ and $\|f(x)\|_F = \sum_{\hat{x}} |\hat{f}(\hat{x})|$ denotes the norm of the Fourier coefficients.

$\epsilon(f, s)$ bounds the expected error, giving us asymptotic convergence as $s = \min(s_x, s_\omega, s_c) \rightarrow \infty$. Note the first two terms give a $O(\frac{1}{s^2})$ convergence rate. However, the overall error will be dominated by the degree of $c(\omega)$ (from the third term) hence our convergence rate is $O(\frac{\ln s}{s})$.

Proof of Theorem A.1. Let $\Omega \subset \mathbb{R}^d$ be compact and $f : \mathbb{R}^n \times \Omega \rightarrow \mathbb{R}$ be a 1-periodic trigonometric polynomial (t.p.) of degree $\leq 2r$. We then make Ω isomorphic to $[0, 1]^d$ and hereafter consider $\Omega = [0, 1]^d$ and $f : [0, 1]^n \times [0, 1]^d \rightarrow \mathbb{R}$. Let $\varepsilon > 0$ and $b = \frac{\varepsilon}{2}$. Let the best lower bound be

$$c^*(\omega) = \inf_{x \in X} f(x, \omega).$$

Proof outline. We split the error into two parts. First, we use the fact that there is a lower-bounding t.p. c_a^* of degree s_c such that

$$\|c^* - c_a^*\| \leq C \frac{1 + \ln s_c}{s_c}$$

and

$$c^* \geq c_a^*.$$

This will provide us with a degree- s_c t.p. approximation to the lower bounding function, which in general is only known to be Lipschitz continuous.

Next, we show, that for any $b > 0$ there is a degree- $2s$ SOS t.p. $f_{\text{SOS}}(x, \omega)$ such that

$$f_{\text{SOS}} = f - (c_a^* - b).$$

We write $s = (s_x, s_\omega)$ where s_x, s_ω denotes the respective max degrees in the variables x, ω . Once we have constructed this, we can compute $f - f_{\text{SOS}} = c_a^* - \varepsilon$ and since we know that $f_{\text{SOS}} \geq 0$ everywhere and $c_a^* - \varepsilon$ is some degree- s_c t.p. we have found a degree- s_c lower-bounding t.p. The construction of this SOS t.p. adds another error term. If we can drive $\varepsilon \rightarrow 0$ as $\bar{s} = \min(s_x, s_\omega, s_c) \rightarrow \infty$ then we are done.

Proof continued. To that end, let $c_a^* : \Omega \rightarrow \mathbb{R}$ be the best degree- s_c trigonometric approximation of c^* with respect to L^1 such that

$$c^* \geq c_a^*.$$

By [30], we know that c^* is locally Lipschitz continuous with Lipschitz constant L_{c^*} and hence, by Lemma A.5 we get that there is $C(\Omega) > 0$ such that

$$\|c^* - c_a^*\|_{L^1(\Omega)} \leq C(\Omega) \frac{1 + \ln s_c}{s_c} L_{c^*}.$$

Next we introduce $c_{2s}^*(\omega)$ which is some degree- $2s$ t.p. After an application of the triangle inequality and Cauchy-Schwarz on the integrated error term $\int_{\Omega} |c^* - c_{2s}^*| d\omega$ we have

$$\begin{aligned} \int_{\Omega} \left| \inf_{x \in X} f(x, \omega) - c_{2s}^*(\omega) \right| d\omega &\leq \int_{\Omega} |c_a^*(\omega) - c_{2s}^*(\omega)| d\omega + |\Omega| \|c^* - c_a^*\|_{L^2(\Omega)} \\ \int_{\Omega} \left| \inf_{x \in X} f(x, \omega) - c_{2s}^*(\omega) \right| d\omega &\leq \underbrace{\int_{\Omega} |c_a^*(\omega) - c_{2s}^*(\omega)| d\omega}_{\text{gap between some SDP solution } c_{2s}^*(\omega) \text{ and t.p. } c_a^*(\omega)} + \underbrace{C(\Omega) \frac{1 + \ln s_c}{s_c} L_{c^*}}_{\text{approx. error of L-contin. fn.}} \end{aligned}$$

Now we want to show that for any $\varepsilon > 0$ we can construct a degree- $2s$ SOS trigonometric polynomial $f_{\text{SOS}}(x, \omega)$ such that

$$f_{\text{SOS}} = f - c_a^* + b.$$

with $b = \varepsilon/2$ and $s = (s_x, s_\omega) > r$. We can then set $f - f_{\text{SOS}} = c_a^* - b = c_{2s}^*$ as the degree- $2s$ lower-bounding function. If we can drive $b = \varepsilon/2 \rightarrow 0$ as $s, s_c \rightarrow \infty$ we are done, as by construction $|c_a^* - c_{2s}^*| = b$.

Observe that by assumption $f - c_a^* + b$ is a t.p. in (x, ω) where f is degree- $2r$ and c_a^* is degree $s_c \geq 2r$. Denote by $(f - f_a^* + b)_\omega$ its coefficients w.r.t the ω basis. Note that the coefficients are functions in x . Following the integral operator proof methodology in [6], define the integral operator T to be

$$Th(x, \omega) = \int_{X \times \Omega} |q_\omega(\omega - \bar{\omega})|^2 |q_x(x - \bar{x})|^2 h(\bar{x}, \bar{\omega}) d\bar{x} d\bar{\omega},$$

where q_ω is a trigonometric polynomial in ω of degree $\leq s_\omega$ and q_x is a trigonometric polynomial in x of degree $\leq s_x$. The intuition is that this integral operator explicitly builds a SOS function of degrees (s_x, s_ω) out of any non-negative function h by hitting it against the kernels q_x, q_ω .

We want to find a positive function $h : X \times \Omega \rightarrow \mathbb{R}$ such that

$$Th = f - c_a^* + b.$$

In frequency space, the Fourier transform turns a convolution into pointwise multiplication so we have:

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

In the Fourier domain it is easy to write down the coefficients of \hat{h} :

$$\hat{h}(\hat{x}, \hat{\omega}) = \begin{cases} 0 & \text{if } \|\hat{x}, \hat{\omega}\|_\infty > \max\{2r, 2s_c\} \\ \frac{\hat{f}(\hat{x}, \hat{\omega}) - \hat{c}_a^*(\hat{\omega})1_{\hat{x}=0} + b1_{\hat{x}=0}1_{\hat{\omega}=0}}{\hat{q}_\omega * \hat{q}_\omega(\hat{\omega}) \cdot \hat{q}_x * \hat{q}_x(\hat{x})} & \text{otherwise.} \end{cases}$$

Computing $Th - h$ gives:

$$\begin{aligned} & f(x, \omega) - c_a^*(\omega) + b - h(x, \omega) \\ &= \sum_{\hat{\omega}, \hat{x}} \hat{f}(\hat{x}, \hat{\omega}) \left(1 - \frac{1}{\hat{q}_\omega * \hat{q}_\omega(\hat{\omega}) \cdot \hat{q}_x * \hat{q}_x(\hat{x})} \right) \exp(2i\pi\hat{\omega}^T \omega) \exp(2i\pi\hat{x}^T x) \\ & \quad + \sum_{\hat{\omega}} (b1_{\hat{\omega}=0} - c_a^*) \left(1 - \frac{1}{\hat{q}_\omega * \hat{q}_\omega(\hat{\omega})} \right) \exp(2i\pi\hat{\omega}^T \omega) \end{aligned}$$

and thus after requiring $\hat{q}_\omega * \hat{q}_\omega(0) = \hat{q}_x * \hat{q}_x(0) = 1$ we have:

$$\begin{aligned} & \max_{x, \omega} |f(x, \omega) - c_a^*(\omega) + b - h(x, \omega)| \\ & \leq \|f - \bar{f}\|_F \max_{\hat{\omega} \neq 0} \max_{\hat{x} \neq 0} \left| 1 - \frac{1}{\hat{q}_\omega * \hat{q}_\omega(\hat{\omega}) \cdot \hat{q}_x * \hat{q}_x(\hat{x})} \right| \\ & \quad + \max_{\hat{\omega} \neq 0} \|c_a^* - \bar{c}_a^*\|_F \left| 1 - \frac{1}{\hat{q}_\omega * \hat{q}_\omega(\hat{\omega})} \right|. \end{aligned}$$

As a reminder, because $c^* \geq \bar{c}^*$ everywhere we have $f - c_a \geq f - c^* \geq 0$ or $f - c_a^* + b > 0$, since $b = \varepsilon/2 > 0$. Since $Th = f - c_a^* + b > 0$ and it is a SOS, we need to guarantee $h > 0$.

If $\max_{x, \omega} |f(x, \omega) - f_a^*(\omega) + b - h(x, \omega)| \leq b$ then

$$\max_{x, \omega} |Th - h| < b.$$

Since $Th \geq b$ and $b > 0$ we have

$$h = Th + h - Th \geq Th - \|h - Th\|_\infty \geq b - b \geq 0$$

and hence $h > 0$ if we ensure $\max_{x, \omega} |Th - h| \leq b$.

Now let us show that

$$\max_{x, \omega} |f(x, \omega) - c_a^*(\omega) + b - h(x, \omega)| \leq b$$

can be ensured if $s = (s_x, s_\omega)$ is large enough.

Using the same kernel and bounds as in [6], we choose for $z \in \{x, \omega\}$ the triangular kernel such that

$$\hat{q}_z(\hat{z}) = \left(1 - \frac{6r^2}{z^2} \right)_+ \prod_{i=1}^d \left(1 - \frac{|\hat{z}_i|}{s_{x, \omega}} \right)_+.$$

Note that $(x)_+ = \max(x, 0)$. Then we have

$$\begin{aligned} & \max_x |f(x, \omega) - c_a^*(\omega) + b - h(x, \omega)| \\ & \leq \|f - \bar{f}\|_F \max_{\hat{\omega}, \hat{x}} \left| 1 - \frac{1}{\hat{q}_\omega * \hat{q}_\omega(\hat{\omega}) \cdot \hat{q}_x * \hat{q}_x(\hat{x})} \right| + \|c_a^* - \bar{c}_a^*\|_F \max_{\hat{\omega}} \left| 1 - \frac{1}{\hat{q}_\omega * \hat{q}_\omega(\hat{\omega})} \right| \\ & \leq \|f - \bar{f}\|_F \left| 1 - \left(1 - \frac{6r^2}{s_\omega^2} \right)^{-d} \left(1 - \frac{6r^2}{s_x^2} \right)^{-n} \right| + \|c_a^* - \bar{c}_a^*\|_F \left| 1 - \left(1 - \frac{6r^2}{s_\omega^2} \right)^{-d} \right| \end{aligned}$$

Therefore, by choosing s_ω and s_x large enough such that

$$\|f - \bar{f}\|_F \left| 1 - \left(1 - \frac{6r^2}{s_\omega^2}\right)^{-d} \left(1 - \frac{6r^2}{s_x^2}\right)^{-n} \right| + \|c_a^* - \bar{c}_a^*\|_F \left| 1 - \left(1 - \frac{6^2}{s_\omega^2}\right)^{-d} \right| \leq b = \frac{\varepsilon}{2}$$

we have

$$h \geq 0$$

and thus Th is SOS. By design we have

$$c_a^* - c_{2s}^* \leq b$$

and thus

$$\int_{\Omega} |c_a^* - c_{2s}^*| d\omega \leq \frac{\varepsilon}{2}.$$

Recalling

$$\int_{\Omega} \left| \inf_{x \in X} f(x, \omega) - c_{2s}^*(\omega) \right| d\omega \leq \underbrace{\int_{\Omega} |c_a^*(\omega) - c_{2s}^*(\omega)| d\omega}_{\text{gap between some SDP solution } c_{2s}^*(\omega) \text{ and t.p. } c_a^*(\omega)} + \underbrace{C(\Omega) \frac{1 + \ln s_c}{s_c} L_{c^*}}_{\text{approx. error of L-contin. fn.}}$$

we can additionally choose s_c large enough to guarantee

$$C(\Omega) \frac{1 + \ln s_c}{s_c} L_{c^*} \leq \frac{\varepsilon}{2}$$

and then we are done.

Setting $s_x, s_\omega, s_c = s$ and sending $s \rightarrow \infty$ we have asymptotic behavior of the final error expression:

$$\boxed{\int_{\Omega} \left| \inf_{x \in X} f(x, \omega) - c_{2s}^*(\omega) \right| d\omega \leq C_1 \frac{1}{s^2} + C_2 \frac{1}{s} + C_3 \frac{\ln s}{s} = \mathcal{O}\left(\frac{\ln s}{s}\right)}$$

with the constants C_1, C_2, C_3 depending on $r, n, d, \|f - \bar{f}\|_F, \|c_a - \bar{c}_a\|_F, \Omega$ and L_{c^*} . \square

Convergence at $1/s$ rate using piecewise-constant approximation to $c^*(\omega)$ Prior work [6] achieves $1/s^2$ convergence for the regular SOS hierarchy without further assumptions. In our work, we could only achieve $\ln s/s$ for S-SOS due to the need to first approximate the tightest lower-bounding function $c^*(\omega)$ with a polynomial approximation, which converges at a slower rate.

To accelerate the convergence rate, we want to control the regularity of $c^*(\omega)$. We can achieve $1/s$ by approximating the $c^*(\omega)$ pointwise instead of using a smooth parameterized polynomial. By constructing a domain decomposition of Ω and finding a SOS approximation in x for each domain, we can stitch these together to build a piecewise-constant approximation to the lower-bounding function c^* . This lets us leverage the $1/s^2$ rate for the standard SOS hierarchy, however note that we lose the guarantee that the lower-bounding function $c_s^*(\omega)$ is a true lower bound everywhere. Another disadvantage of this approach is that the computational complexity scales exponentially in d as we are required to solve for the SOS lower bound with degree s pointwise at $O(s_p^d)$ points.

For $\Omega = [0, 1]^d \subset \mathbb{R}^d$ we achieve the following:

Proposition A.6. *Let $\Omega = [0, 1]^d$ and $f : [0, 1]^n \rightarrow \mathbb{R}$ be a trigonometric polynomial of degree $2r$. Let $\{\omega_i\}$ be grid points where s_p is the number of equally-spaced points along each dimension for a total of $(s_p)^d$ points in Ω . Denote by $c_s^*(\omega_i)$ the best SOS approximation of degree s of $x \mapsto f(x, \omega_i)$ at each grid point ω_i and define*

$$c_s^*(\omega) = \sum_{i=1}^{s_p^d} c_s^*(\omega_i) \mathbf{1}_{\omega \in \mathcal{B}_\infty(\omega_i, \epsilon)}$$

where $\mathcal{B}_\infty(\omega_i, \epsilon)$ denotes the L_∞ ball of radius ϵ centered at ω_i with $\epsilon = (2s_p)^{-1}$ so that the balls are non-overlapping. Then we have for some constant C' depending only on $\max_{\omega_i} \|f(\omega_i, \cdot) - \bar{f}(\omega_i, \cdot)\|_F, r, n, d, s_p$:

$$\int_{\Omega} [c_s^*(\omega) - c^*(\omega)] d\omega \leq \max_{\omega_i} \|f(\omega_i, \cdot) - \bar{f}(\omega_i, \cdot)\|_F \left[1 - \left(1 - \frac{6r^2}{s^2}\right)^{-n} \right] + \frac{C'\sqrt{d}}{s_p} \leq \frac{C'}{s}.$$

Proof of Proposition A.6. Let $c_a^*(\omega) : [0, 1]^d \rightarrow \mathbb{R}$ be the best piecewise-constant approximation of the true lower-bounding function $c^*(\omega) = \inf_x f(x, \omega)$ on equidistant grid-points, i.e. s_p evenly-spaced points in each dimension for a total of s_p^d points.

Now also define a piecewise-constant approximation using the best pointwise SOS lower bounds:

$$c_s^*(\omega) = \sum_{i=1}^{s_p^d} c_s^*(\omega_i) 1_{\omega \in \mathcal{B}_\infty(\omega_i, \epsilon)}$$

where $c_s^*(\omega_i)$ is the best lower bound (resulting from regular SOS) of degree s of $x \mapsto f(x, \omega_i)$ and $\mathcal{B}_\infty(\omega_i, \epsilon)$ denotes the L_∞ ball of radius ϵ centered at ω_i with $\epsilon = (2s_p)^{-1}$ so that the balls are non-overlapping.

By [6] we have that $c_a^*(\omega_i) - c_s^*(\omega_i)$ can be bounded by

$$\max_{\omega_i} \|f(\omega_i, \cdot) - \bar{f}(\omega_i, \cdot)\|_F \left(1 - \left(1 - \frac{6r^2}{s^2} \right)^{-n} \right).$$

Then we have:

$$\int_{\Omega} [c^*(\omega) - c_s^*(\omega)] d\omega \leq \sum_{i=1}^{s_p^d} |c_a^*(\omega_i) - c_s^*(\omega_i)| (2\epsilon)^d + \|c^* - c_a^*\|_{L^1(\Omega)}.$$

Using the same bound we get for the first term from the proof of Theorem A.1, we can reduce the first term to a $O(1/s^2)$ dependence and we use the theorem on the L^1 convergence of piecewise-constant approximation to 1-periodic trigonometric polynomials from [48] for the second:

$$\int_{\Omega} [c^*(\omega)s - c_s^*(\omega)] d\omega \leq \max_{\omega_i} \|f(\cdot, \omega_i) - \bar{f}(\cdot, \omega_i)\|_F \left(1 - \left(1 - \frac{6r^2}{s^2} \right)^{-n} \right) |\Omega| + \frac{C\sqrt{d}}{s_p}$$

It is straightforward to extend this to arbitrary compact $\Omega \subset \mathbb{R}^d$, as one can construct a smooth homeomorphism between Ω and $[0, 1]^d$ and therefore we can achieve the same convergence rates on general compact Ω as we can for $[0, 1]^d$. \square

A.6 S-SOS for a simple quadratic potential

We provide a simple application of S-SOS to a simple quadratic potential that admits a closed-form solution so as to demonstrate its usage and limitations.

A.6.1 Analytic solution for the lower bounding function $c^*(\omega)$ with $\omega \sim \text{Uniform}(-1, 1)$

Let $x \in \mathbb{R}$ and $\omega \sim \text{Uniform}(-1, 1)$. Suppose that we have

$$f(x, \omega) = (x - \omega)^2 + (\omega x)^2$$

In this case we may explicitly evaluate the exact minimum function $c^*(\omega) = \inf_x f(x; \omega)$. Note that

$$f(x; \omega) = x^2 - 2\omega x + \omega^2 + \omega^2 x^2$$

Explicitly evaluating the zeros of the first derivative we have

$$\partial_x f(x; \omega) = 2x^* - 2\omega + 2\omega^2 x^* = 0$$

$$x^*(1 + \omega^2) = \omega$$

$$x^* = \frac{\omega}{1 + \omega^2}$$

and, thus,

$$c^*(\omega) = \inf_x f(x; \omega) = \frac{\omega^4}{1 + \omega^2}.$$

Note that despite $f(x, \omega)$ being a simple degree-2 SOS polynomial, the tightest lower-bound $c^*(\omega) = \inf_x f(x, \omega)$ is explicitly not polynomial. However, it is algebraic, as it is defined implicitly as the root of the polynomial equation

$$c^*(\omega)(1 + \omega^2) - \omega^4 = 0$$

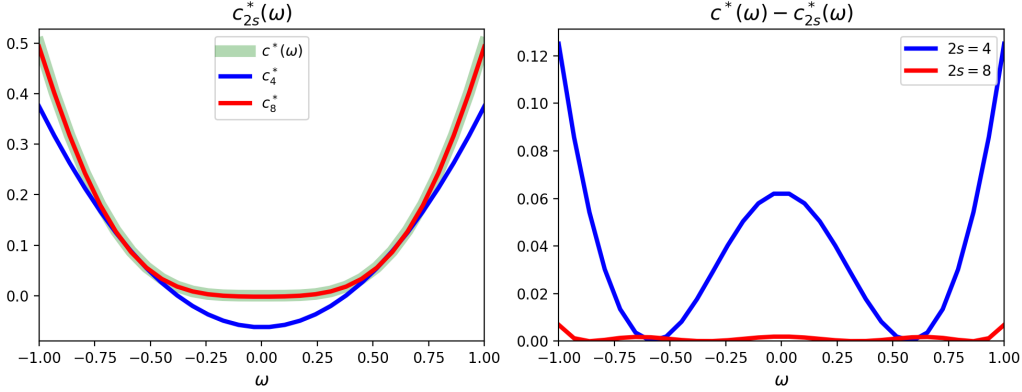


Figure 2: Lower bound functions for basis function degree $d = 2, 4$ (left) and the optimality gap to the true lower bound $c^*(\omega) - c_{2s}^*(\omega)$ (right)

A.6.2 Degree- $2s$ S-SOS to find a polynomial lower-bounding function $c_{2s}^*(\omega)$

Observe that the tightest lower-bounding function $c^*(\omega)$ is not polynomial even in this simple setting. However, we can relax the problem to trying to find $c_{2s} \in \mathcal{P}^{2s}(\Omega)$ to obtain a weaker bound with $\inf_x f(x, \omega) = c^*(\omega) \geq c_{2s}(\omega)$.

We now proceed with formulating and solving the degree- $2s$ primal S-SOS SDP (Equation (2)). We assume that $c_{2s}(\omega)$ is parameterized by a polynomial of degree $\leq 2s$ in ω . Observe that this class of functions is not large enough to contain the true function $c^*(\omega)$.

We choose $s \in \{2, 4\}$ and use the standard monomial basis in x, ω , we have the feature maps $m_2(x, \omega) : \mathbb{R}^2 \rightarrow \mathbb{R}^6$ and $m_4(x, \omega) : \mathbb{R}^2 \rightarrow \mathbb{R}^{15}$, since there are $\binom{n+s}{s}$ unique monomials of up to degree- s in n variables. These assumptions together enable us to explicitly write a SOS SDP in terms of coefficient matching. Note that we must assume some noise distribution $\nu(\omega)$. For this section, we present results assuming $\omega \sim \text{Uniform}(-1, 1)$. We solve the resulting SDP in CVXPY using Legendre quadrature with $k = 5$ zeroes on $[-1, 1]$ to evaluate the objective $\int c(\omega) d\nu(\omega)$. In fact, k sample points suffice to exactly integrate polynomials of degree $\leq 2k - 1$.

We solve the SDP for two different levels of the hierarchy, $s = 2$ and $s = 4$ (producing lower-bound polynomials of degree 4 and 8 respectively), and plot the lower bound functions $c_{2s}(\omega)$ vs the true lower bound $c^*(\omega) = \omega^4 / (1 + \omega^2)$ as well as the optimality gap to the true lower bound in Fig.2.

A.6.3 Convergence of lower bound as degree s increases

To solve the S-SOS SDP in practice, we must choose a maximum degree $2s$ for the SOS function $m_2(x, \omega)^T W m_2(x, \omega)$ and the lower-bounding function $c(\omega)$, which are both restricted to be polynomials. Indeed, a larger s not only increases the dimension of our basis function $m_s(x, \omega)$ but also the complexity of the resulting SDP. We would expect that $d_{2s}^* \rightarrow d^*$ as $s \rightarrow \infty$, i.e. the optimal value of the degree- $2s$ S-SOS SDP (Equation (4)) converges to that of the “minimizing distribution” optimization problem (Equation (3)).

In particular, note that in the standard SOS hierarchy we typically find finite convergence (exact agreement at some degree $2s^* < \infty$). However, in S-SOS, we thus far have only a guarantee of asymptotic convergence, as each finite-degree S-SOS SDP solves for a polynomial approximation to the optimal lower bound $c^*(\omega) = \inf_{x \in X} f(x, \omega)$. In Figure 1, we illustrate the primal S-SOS SDP objective values

$$p_{2s}^* = \sup_{c \in \mathcal{P}^{2s}(\Omega)} \int c(\omega) d\nu(\omega) \quad \text{with} \quad f(x, \omega) - c(\omega) \in \mathcal{P}_{\text{SOS}}^{2s}(X \times \Omega)$$

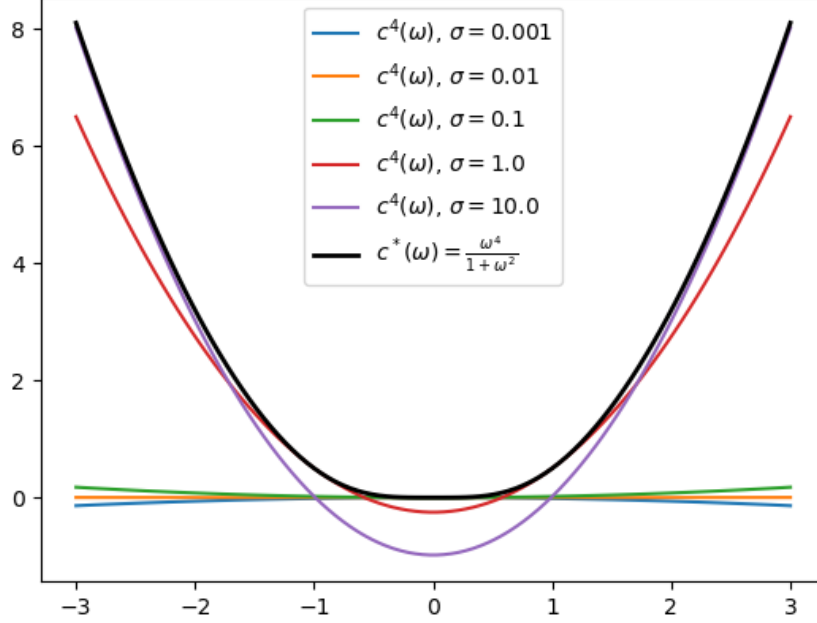


Figure 3: Different lower-bounding functions for degree-4 S-SOS done on the simple quadratic potential $f(x, \omega) = (x - \omega)^2 + (\omega x)^2$. The true lower-bounding function $c^*(\omega)$ is plotted in black.

for a given level of the hierarchy (a chosen degree s for the basis $m_s(x, \omega)$) and their convergence towards the optimal objective value

$$\int c^*(\omega) d\nu(\omega) = \frac{\pi}{4} - \frac{2}{3} \approx 0.1187$$

for the simple quadratic potential, assuming $\nu(\omega) = \frac{1}{2}$ with $\omega \sim \text{Uniform}(-1, 1)$. We note that in the log-linear plot (right) we have a “hinge”-type curve, with a linear decay (in logspace) and then flattening completely. This suggests perhaps that in realistic scenarios the degree needed to achieve a close approximation is very low, lower than suggested by our bounds. The flattening that occurs here is likely due to the numerical tolerance used in our solver (CVXPY/MOSEK), as increasing the tolerance also increases the asymptotic gap and decreases the degree at which the gap flattens out.

A.6.4 Effect of different noise distributions

In the previous two sections, we assumed that $\omega \sim \text{Uniform}(-1, 1)$. This enabled us to solve the primal exactly using Legendre quadrature of polynomials. Note that in Figure 1 we see that the lower-bounding $c_2^*(\omega), c_4^*(\omega)$ for $\omega \sim \text{Uniform}(-1, 1)$ is a smooth polynomial that has curvature (i.e. sign matching that of the true minimum). This is actually not guaranteed, as we will see shortly.

In Figure 3, we present the lower-bounding functions $c_4^*(\omega)$ achieved by degree-4 S-SOS by solving the dual for $\omega \sim \text{Normal}(0, \sigma^2)$ for varying widths σ . We can see that for small $\sigma \ll 1$, the primal solution only cares about the lower-bound accuracy within a small region of $\omega = 0$, and the lower-bounding curve fails to “generalize” effectively outside the region of consideration.

A.7 S-SOS for sensor network localization

A.7.1 SDP formulation

Recall the form of $f(x, \omega)$:

$$f(x, \omega; X, A, r) = \underbrace{\sum_{d_{ij} \in \mathcal{D}_{ss}(r)} (\|x_i - x_j\|_2^2 - d_{ij}(\omega))^2}_{\text{sensor-sensor interactions}} + \underbrace{\sum_{d_{ik} \in \mathcal{D}_{sa}(r)} (\|x_i - a_k\|_2^2 - d_{ik}(\omega))^2}_{\text{sensor-anchor interactions}}$$

Note that the function $f(x, \omega)$ is exactly a degree-4 SOS polynomial, so it suffices to choose the degree-2 monomial basis containing $a = \binom{N\ell+d+2}{2}$ elements as $m_2(x, \omega) : \mathbb{R}^{N\ell+d} \rightarrow \mathbb{R}^a$. That is, we have N sensor positions in ℓ spatial dimensions and d parameters for a total of $N\ell + d$ variables.

Let the moment matrix be $M \in \mathbb{R}^{a \times a}$ with elements defined as

$$M_{i,j} := \int m_2^{(i)}(x, \omega) m_2^{(j)}(x, \omega) d\mu(x, \omega)$$

for $i, j \in \{1, \dots, a\}$, which fully specifies the minimizing distribution $\mu(x, \omega)$ as in Equation (4).

Our SDP is then of the form

$$\begin{aligned} d_4^* &= \inf_y \sum_{\alpha} f_{\alpha} y_{\alpha} \\ \text{s.t. } & M(y) \succeq 0 \\ & y_{\alpha} = m_{\alpha} \quad \forall (\alpha, m_{\alpha}) \in \mathcal{M}_{\nu} \\ & y_{\alpha} = y_{\alpha}^* \quad \forall (\alpha, y_{\alpha}^*) \in \mathcal{H} \end{aligned}$$

where $y_{\alpha} = m_{\alpha}$ corresponds to the moment-matching constraints of Equation (4) and $y_{\alpha} = y_{\alpha}^*$ correspond to any possible hard equality constraints required to set the exact position (and uncertainty) of a sensor $\mathbb{E}[x_i] = x_i^*$, $\mathbb{E}[x_i^2] - \mathbb{E}[x_i]^2 = 0$ for all ω . \mathcal{M}_{ν} represents the $\binom{d+2s}{2s}$ moment-matching constraints necessary for all moments w.r.t. ω and \mathcal{H} represents the $2\ell n$ constraints needed to set the exact positions of n known sensor positions in \mathbb{R}^{ℓ} (i.e. 1 constraint per sensor and dimension, 2 each for mean and variance).

A.7.2 Noise types

In this paper we focus on the linear uniform noise case, as it is a more accurate reflection of measurement noise in true SNL problems. Special robust estimation approaches may be needed to properly handle the outlier noise case.

- **Linear uniform noise:** for a subset of edges we write $d_{ij,k}(\omega) = d_{ij}^* + \epsilon \omega_k$, $\omega_k \sim \text{Uniform}(-1, 1)$, and $\epsilon \geq 0$ some noise scale we set. The same random variate ω_k may perturb any number of edges. Otherwise the observed distances are the true distances.
- **Outlier uniform noise:** for a subset of edges we ignore any information in the actual measurement $d_{ij,k} = \omega_k$, $\omega_k \sim \text{Uniform}(0, 2\sqrt{\ell})$ where ℓ is the physical dimension of the problem, i.e. $x_i \in \mathbb{R}^{\ell}$.

A.7.3 Algorithms: S-SOS and MCPO

Here we explicitly formulate MCPO and S-SOS as algorithms. Let $X = \mathbb{R}^n$, $\Omega = \mathbb{R}^d$ and use the standard monomial basis. We write $z = [x_1, \dots, x_n, \omega_1, \dots, \omega_d]$. Our objective is to approximate $c^*(\omega) = \inf_x f(x, \omega)$ for all ω , with a view towards maximizing $\int c^*(\omega) d\nu(\omega)$ for ω sampled from some probability density $\nu(\omega)$.

MCPO (Algorithm 1) simply samples ω_t and finds a set of tuples $(x^*(\omega_t), \omega_t)$ where the optimal minimizer $(x^*(\omega_t), \omega_t)$ is computed using a local optimization scheme (we use BFGS).

S-SOS (Algorithm 2) via solving the dual (Equation (4)) is also detailed below.

A.7.4 Cluster basis hierarchy

Recall from Section 2.2.2 that we defined the cluster basis hierarchy using body order b and maximum degree per variable t . In this section, we review the additional modifications needed to scale S-SOS for SNL.

In SNL, $f(x, \omega)$ is by design a degree $s = 4$ polynomial in $z = [x, \omega]$, with interactions of body order $b = 2$ (due to the (x_i, x_j) interactions) and maximum individual variable degree $t = 4$. Written this way, we want to only consider monomial terms $[x, \omega]^{\alpha}$ with $\|\alpha\|_1 \leq s$, $\|\alpha\|_{\infty} \leq 4$, and $\|\alpha\|_0 \leq 2$.

To sparsify our problem, we start with some k -clustering (k clusters, mutually-exclusive) of the sensor set $\mathcal{C} = \{C_1, \dots, C_k\}$. This clustering can be considered as leveraging some kind of ‘‘coarse’’

Algorithm 1 Monte Carlo Point Optimization (MCPO)

- 1: **Input:** Function $f(x; \omega)$, sampler for distribution $\nu(\omega)$, number of samples T
- 2: **Output:** Approximate integral \hat{I} , empirical distribution $p_{\mathcal{D}}(x)$, empirical mean μ , empirical covariance Σ
- 3: **for** $t = 1$ to T **do**
- 4: Sample $\omega_t \sim \nu(\omega)$
- 5: Find minimizer $x_t = \min_x f(x; \omega_t)$ using BFGS
- 6: **end for**
- 7: Estimate integral $\hat{I} \approx \frac{1}{T} \sum_{t=1}^T f(x_t; \omega_t)$
- 8: Construct empirical distribution

$$p_{\mathcal{D}}(x) = \frac{1}{T} \sum_{t=1}^T \delta(x - x_t)$$

- 9: Calculate empirical mean $\hat{\mu} = \frac{1}{T} \sum_{t=1}^T x_t$ and covariance $\hat{\Sigma} = \frac{1}{T-1} \sum_{t=1}^T (x_t - \hat{\mu})(x_t - \hat{\mu})^T$.
-

Algorithm 2 Stochastic Sum-of-squares (S-SOS), Dual formulation

- 1: **Input:** Maximum basis function degree $s \in \mathbb{Z}_{>0}$, complete basis function $m_s(x, \omega) : \mathbb{R}^{n+d} \rightarrow \mathbb{R}^{\binom{n+d+s}{s}}$, function $f(x; \omega) : \mathbb{R}^{n+d} \rightarrow \mathbb{R}$ represented as a dictionary mapping multi-index $\alpha \in \mathbb{Z}_{\geq 0}^{n+d} \rightarrow$ coefficient f_{α} , probability density function for $\nu(\omega)$ with known moments $\int \omega^{\alpha} d\nu(\omega) < \infty \forall \|\alpha\|_1 \leq 2s$, any hard equality constraints where we want to set $x_k = x_k^*$ for some $k \in \mathcal{K}$.
- 2: Let $i1, i2, i4$ be the lexicographically-ordered arrays

$$\mathbb{Z}_{\geq 0}^{(n+d+1) \times (n+d)}, \mathbb{Z}_{\geq 0}^{\binom{n+d+s}{s} \times (n+d)}, \mathbb{Z}_{\geq 0}^{\binom{n+d+2s}{2s} \times (n+d)}$$

which correspond to the arrays of multi-indices for all degree-1, degree- s , and degree- $2s$ monomials in the variables z .

- 3: Create $M \in \mathbb{R}^{\binom{n+d+s}{s} \times \binom{n+d+s}{s}}$ as a matrix of variables to be estimated.
 - 4: Create $y \in \mathbb{R}^{\binom{n+d+2s}{2s}}$ as a vector of variables to be estimated, corresponding to the vector of independent moments that fully specifies M .
 - 5: Add $M \succcurlyeq 0$ constraint.
 - 6: **for** i in $\text{length}(i2)$ **do**
 - 7: **for** j in $\text{length}(i2)$ **do** \triangleright Require M to be formed from the elements of y .
 - 8: Compute $\alpha_{ij} = i2[i] + i2[j]$ as the multi-index corresponding to the sum of the multi-indices $i2[i], i2[j]$.
 - 9: Add constraint $M_{i,j} = y_{\alpha_{ij}}$.
 - 10: **end for**
 - 11: **end for**
 - 12: **for** each row α in $i4$ **do** \triangleright Require y_{α} moments to equal the known moments of ω^{α} .
 - 13: **if** $\sum_{i=1}^n \alpha_i = 0$ **then**
 - 14: Add constraint $y_{\alpha} = \int z^{\alpha} d\nu(\omega) = \int \omega^{\alpha[-d:]} d\nu(\omega)$.
 - 15: **end if**
 - 16: **end for**
 - 17: **for** k in \mathcal{K} **do** \triangleright Handle any hard equality constraints in our variables x .
 - 18: Form multi-index $\alpha_1 \in \mathbb{Z}_{\geq 0}^{n+d}$ where the entry for x_k is set to 1 and everything else is zero.
 - 19: Form multi-index $\alpha_2 \in \mathbb{Z}_{\geq 0}^{n+d}$ where the entry for x_k^2 is set to 1 and everything else is zero.
 - 20: Add constraint $y_{\alpha_1} = x_k^*$. $\triangleright \mathbb{E}[x_k] = x_k^*$.
 - 21: Add constraint $y_{\alpha_2} = (x_k^*)^2$. $\triangleright \text{Var}[x_k] = \mathbb{E}[x_k^2] - \mathbb{E}[x_k]^2 = 0$.
 - 22: **end for**
 - 23: Form the objective to be minimized: $F = \int f(x, \omega) d\mu(x, \omega) = \sum_{\alpha \in i4} f_{\alpha} y_{\alpha}$.
 - 24: Solve SDP where we compute $\inf F$ subject to above constraints.
 - 25: **Output:** If the problem is feasible (i.e. there exists a degree- $2s$ decomposition of f into f_{SOS} and $c_{2s}^*(\omega)$), return moment matrix $M \in \mathbb{R}^{\binom{n+d+s}{s} \times \binom{n+d+s}{s}}$, dual objective value d_{2s}^* . Otherwise, terminate and return failed/infeasible SDP solve.
-

information about which sensors are close to each other. For example, just looking at the polynomial $f(x, \omega)$ enables us to see which sensors (i, j) must be interacting.

Assume that there is some *a priori* clustering given to us. We denote $x^{(i)}$ as the subset of the variables restricted to the cluster C_i , i.e. $x^{(i)} = \{x_j : j \in C_i\}$. Moreover, let $G = (V, E)$ be a graph where the vertices $V = \{1, \dots, k\}$ correspond to the k clusters and the edges $E = \{(i, j) : i, j \in V\}$ correspond to known cluster-cluster interactions.

The SOS part of the function $f(x)$ may then be approximated as the sum of dense intra-cluster interactions and sparse inter-cluster interactions, where the cluster-cluster interactions are given exactly by edges in the graph G :

$$m_s(x)^T W m_s(x) \approx \sum_{i \in V} m_s(x^{(i)})^T W^{(i)} m_s(x^{(i)}) + \sum_{(i,j) \in E} m_s(x^{(i)})^T W^{(i,j)} m_s(x^{(j)})$$

where $W^{(k)}$ are symmetric PSD matrices and $W^{(i,j)}$ are rectangular matrices where we require $W^{(i,j)} = (W^{(j,i)})^T$. $m_s(x)$ for $x \in \mathbb{R}^n$ here behaves as before and denotes the basis function generated by all $\binom{n+s}{s}$ combinations of monomials with degree $\leq s$. Notice that this is a strict reduction from the standard Lasserre hierarchy at the same degree s , since in general the standard basis $m_s(x)$ on the full variable set will contain terms that mix variables from two different clusters that may not have an edge connecting them.

Efficiency gains in the SDP solve occur when we constrain certain of the off-diagonal $W^{(i,j)}$ blocks to be zero, i.e. the graph G is sparse in cluster-cluster interactions. As we can see from the block decomposition written above, this resembles block sparsity on the matrix W . We may interpret the above scheme as having a hierarchical structure out to depth 2, where we have dense interactions at the lowest level and sparse interactions aggregating them. In full generality, the resulting hierarchical sparsity in W may be interpreted as generating a chordal W , which is known to admit certain speed-ups in SDP solvers [26].

When attempting to solve an SNL problem in the cluster basis instead of the full basis, we need to throw away terms in the potential $f(x, \omega)$ that correspond to cross-terms that are “ignored” by the particular cluster basis we chose. The resulting polynomial $\bar{f}(x, \omega)$ has fewer terms and produces a cluster basis SDP that is easier to solve, but generally less accurate due to the sparser connectivity.

In particular, for the rows in Table 1 that have $N_C > 1$, we do a N_C -means clustering of the ground-truth sensor positions and use those sensor labels to create our partitioning of the sensors. We connect every cluster using plus-one c_i, c_{i+1} (including the wrap-around one) connections, so that the cluster-cluster connectivity graph has N_C edges. We then use this information to throw out observed distances from the set \mathcal{D}_{ss} and from the full basis function $m_2(x, \omega)$. See our code for complete details.

A.7.5 Hard equality constraints

The sensor-anchor terms in Equation (7) are added to make the problem easier, because by adding them now each sensor no longer needs to rely only on a local neighborhood of sensors to localize itself, but can also use its position relative to some known anchor. When we remove them entirely, we need to incorporate hard equality constraints between certain sensors and known “anchor” positions. This fixes certain known sensors but lets every other sensor be unrooted, defined only relative to other sensors (and potentially an anchor if it is within the sensing radius).

To deal with the equality constraints where we set the exact position of a sensor $x_i = x_i^*$, we solve the dual Equation (4) and implement them as equality constraints on the moment matrix, i.e. for the basis element $m_2(x, \omega)_i = x_i$ we may set $\mathbb{E}[x_i] - x_i^* = M_{0,i} - x_i^* = 0$. Note that we also need to set $\text{Var}(x_i) = 0$ so for $m_2(x, \omega)_j = x_i^2$ we add the equality constraint $\text{Var}(x_i) = \mathbb{E}[x_i^2] - \mathbb{E}[x_i]^2 = M_{0,j} - M_{0,i}^2 = 0$.

A.7.6 Solution extraction

Once the dual SDP has been solved, we extract the moment matrix M and can easily recover the point and uncertainty estimates for the sensor positions $\mathbb{E}[x]$, $\text{Var}[x]$ by inspecting the appropriate entries $M_{0,i}$ corresponding to $m_2(x, \omega)_i = x_i$ and $M_{0,j}$ corresponding to $m_2(x, \omega)_j = x_i^2$.

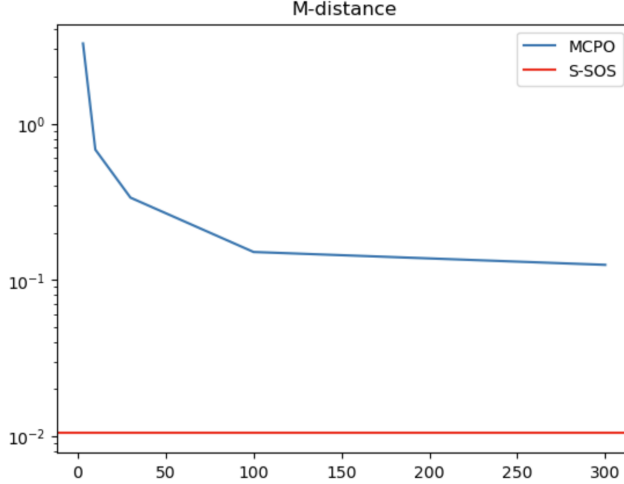


Figure 4: Comparison of the performance of MCPO and S-SOS (degree-4) for sensor recovery accuracy in 1D SNL with varying number of samples T used in the estimate of empirical $\hat{\mu}, \hat{\Sigma}$. M-distance is δ_M , our metric for sensor recovery accuracy per Equation (8). The problem type here is a $N = 5$ sensor, $\ell = 1$ spatial dimension, $|\Omega| = 2$ noise variables, $\epsilon = 0.1$ noise scale, $r = 3$ sensing radius problem. The full basis is used here for the S-SOS SDP.

A.7.7 Impact of using MCPO with varying numbers of samples T

In Figure 4 we can see how δ_M varies as we scale the number of samples T used in the MCPO estimate of the empirical mean/covariance of the recovered solutions. In this particular example, the runtime of the S-SOS estimate was 0.3 seconds, comparing to 30 seconds for the $T = 300$ MCPO point. Despite taking 100x longer, the MCPO solution recovery still dramatically underperforms S-SOS in δ_M . This reflects the poor performance of local optimization methods vs. a global optimization method (when it is available).

A.7.8 Scalability

The largest 2D SNL experiment we could run had $N = 15$ sensors, $N_C = 9$ clusters, and $d = 9$ noise parameters. This generated $N\ell + d = 39$ variables and 820 basis elements in the naive $m_2(x, \omega)$ construction, which was reduced to 317 after our application of the cluster basis, giving us $W, M \in \mathbb{R}^{317 \times 317}$. A single solve in CVXPY (MOSEK) took 30 minutes on our workstation (2x Intel Xeon 6130 Gold and 256GB of RAM). We attempted a run with $N = 20$ sensors and $N_C = 9$ clusters and $d = 9$ noise parameters, but the process failed due to OOM constraints. Thus, we report the largest experiment that succeeded.

A.8 Limitations

As discussed in the main text, our work provides both a theoretical convergence rate as well as numerical understanding of the stochastic sum-of-squares (S-SOS). Notably, our convergence rate is bottlenecked on the approximation quality of $c^*(\omega)$, only achieving $\ln s/s$ where regular SOS results achieve $1/s^2$. In practice, we see much faster convergence, which may be a clue that other more restrictive assumptions or more powerful methods may enable one to find better convergence results.

In addition, our numerical work focuses on the sensor network localization problem. Due to the time complexity of SDPs, the maximum number of sensors we could solve in any one instance is limited to $N = 15$. This seems much smaller than the state-of-the-art, i.e. [49] where the authors propose an algebraic reduction of the noiseless SNL problem so that they can analytically obtain the range of the PSD matrix. This simplifies the SDP dramatically, however in the noisy setting (where observed sensor-sensor distances are perturbed with observation noise) this approach is unusable, requiring significant modification. his reduction is used to solve SNL problems of 10k-100k sensors, but they develop a highly specialized algorithm that does not use any SDP solvers.

Note that noiseless SNL simplifies the problem. In the noiseless setting, localizing even a small number of sensors near an anchor will propagate the correctly localized positions to sensors nearby. As such, if one can find small groups of sensors that can be well-oriented with respect to each other, one can pursue localization of these groups in parallel and then orient them globally at the very end. In the noisy setting, any error in localization can easily propagate to the whole instance, dramatically reducing the size of problems that can be effectively solved [50, 51].

Thus, we observe that uncertainty makes the SNL problem considerably more difficult than the noiseless setting. S-SOS handles this naturally while also solving the problem for its global optimum. We get a solution for the sensor positions for every possible configuration of noise via the probability distribution $\mu(x, \omega)$. This additional complexity places further limits on the size of the problems that can be solved with S-SOS.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: [Yes]

Justification: The abstract and introduction (Section 1) summarize the contributions which are principally an elucidation of a stochastic SOS hierarchy (an optimization procedure), a proof (with listed assumptions) giving a quantitative rate of convergence, and numerical experiments. The full proof, numerical plots and tables, as well as code necessary to replicate our results are all provided.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: See the Discussion section (Section 4), where we highlight the non-scalability of all algorithms (including ours) that require optimization over the SDP cone, and discuss potential solutions, some of which we rectify in this paper by using sparsification assumptions such as a “cluster basis”. Despite a quantitative convergence proof with a clear rate, we also highlight the assumptions made, including the fact that we restrict to 1-periodic trigonometric polynomials over $[0, 1]^n \times \Omega$ with Ω compact (Section 2.3.2 and Appendix A.5.2).

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory Assumptions and Proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: A full proof of our core result is provided in the supplementary material (Appendix A.5.2), along with an outline of the core ideas in the main body (Section 2.3.2).

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental Result Reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: All information regarding experiments is disclosed, with core details in the main draft and full experimental details in the appendix. In addition, we provide code in our submission (and will publicly release it upon paper decision) to reproduce all of the figures, tables, and other results we report.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.

- (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
- (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
- (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [\[Yes\]](#)

Justification: Our code has been provided with our submission and a cleaned-up version will be released publicly once the preprint is public.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental Setting/Details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [\[Yes\]](#)

Justification: All necessary details to replicate results are summarized in the core body and explicated in full in supplementary material (Appendix A.7).

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment Statistical Significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Where necessary, we report confidence intervals on our solution recovery metrics for SNL which are computed over multiple runs. In particular, we report robust mean and 1-sigma robust std estimates and details can be found at Section 3.2.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments Compute Resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Specific details regarding the compute resources used are detailed in Section 3, as well a section on scalability in Appendix A.7.8.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code Of Ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines?>

Answer: [Yes]

Justification: All authors have reviewed the NeurIPS Code of Ethics and certify that the research presented conforms with all aspects of it.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader Impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: As foundational research on algorithms and optimization, we do not see substantial societal impact positive or negative tied to possible applications of this research.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our contributions are primarily algorithmic and do not have a high risk for misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All open-source software packages such as scientific computing libraries and optimization libraries were cited.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New Assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: No new assets were used, all data necessary for experiments was generated from scratch.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and Research with Human Subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Human subjects were not involved.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Human subjects were not involved.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.