Non-negative Polynomial Density Estimation via Sum-of-Squares: Theory, Hard Problems, and the Curse of Dimensionality

AI Research Assistant

November 10, 2025

Abstract

Density estimation is a fundamental problem in statistics, often rendered intractable by complex, non-convex functional shapes and high-dimensional data spaces. While traditional non-parametric methods (like Kernel Density Estimation) struggle with the curse of dimensionality, and parametric methods often lack necessary expressiveness, polynomial approximations offer a powerful alternative. This paper explores the rigorous application of Sum-of-Squares (SOS) optimization to enforce global non-negativity in polynomial density models. We examine the theoretical foundations of realizing the cone of positive polynomials via Semidefinite Programming (SDP). We further detail specific "hard problems" in the literature—such as modeling pathological distributions (e.g., Neal's Funnel) and modern Generative Flows—where SOS techniques provide unique solutions. Finally, we provide a quantitative analysis of the curse of dimensionality as it pertains specifically to the size of the Gram matrices in the resulting SDPs.

1 Introduction

The estimation of an unknown probability density function $f : \mathbb{R}^d \to \mathbb{R}_+$ from independent and identically distributed samples $\{x_i\}_{i=1}^N$ is a cornerstone of statistical inference. The core challenge in approximating f with a flexible function class, such as polynomials, lies in adhering to the fundamental axioms of probability measures:

1. **Non-negativity:** $f(x) \geq 0, \forall x \in \operatorname{supp}(f)$.

2. **Normalization:** $\int_{\operatorname{supp}(f)} f(x)dx = 1$.

While normalization can often be handled via linear constraints on coefficients (for distributions with compact support) or normalizing constants, the global non-negativity constraint is fundamentally difficult.

Checking whether a multivariate polynomial $p(x)$ of degree ≥ 4 is globally non-negative is NP-hard. Consequently, standard optimization techniques cannot directly enforce this constraint during estimation (e.g., Maximum Likelihood Estimation).

The relaxation of non-negativity to the *Sum-of-Squares* (SOS) condition provides a computationally tractable, albeit sufficient rather than necessary, substitute. By leveraging the duality between the SOS cone and the cone of positive semidefinite matrices, density estimation can be cast as a Semidefinite Program (SDP), guaranteeing globally optimal solutions for specific classes of problems.

2 Theoretical Foundations

2.1 The Cone of Non-negative Polynomials vs. SOS

Let $\mathbb{R}[x]_{2d}$ denote the ring of polynomials in n variables with even degree $2d$. We define the cone of non-negative polynomials as:

$$P_{n,2d} = \{p \in \mathbb{R}[x]_{2d} \mid p(x) \geq 0, \forall x \in \mathbb{R}^n\}$$

Optimization over this cone is generally intractable. We instead consider the SOS cone:

$$\Sigma_{\text{SOS}n,2d} = \left\{ p \in \mathbb{R}[x]_{2d} \mid p(x) = \sum_i q_i(x)^2, \text{ for some } q_i \in \mathbb{R}[x]_d \right\}$$

It is clear that $\Sigma_{\text{SOS}n,2d} \subseteq P_{n,2d}$. Hilbert proved in 1888 that these cones are equal only in three distinct cases: univariate polynomials ($n = 1$), quadratic polynomials ($2d = 2$), and bivariate quartics ($n = 2, 2d = 4$). For all other cases, there exist non-negative polynomials that are *not* sums of squares (e.g., the Motzkin polynomial). However, for practical density estimation, Σ_{SOS} is dense enough in the space of continuous non-negative functions to serve as an excellent approximation class.

2.2 SDP Realization

The practical utility of SOS stems from its realization via SDP. A polynomial $p(x)$ of degree $2d$ is SOS if and only if there exists a positive semidefinite matrix $Q \succeq 0$ (the Gram matrix) such that:

$$p(x) = v_d(x)^T Q v_d(x) = \text{Tr}(Q \cdot v_d(x)v_d(x)^T) \tag{1}$$

where $v_d(x)$ is the vector of all monomials up to degree d:

$$v_d(x) = [1, x_1, \ldots, x_n, x_1^2, x_1 x_2, \ldots, x_n^d]^T$$

The dimension of this vector, $N = \binom{n+d}{d}$, determines the size of the SDP.

2.3 Maximum Likelihood Formulation

For Density Estimation, we typically seek to maximize the log-likelihood of the samples. If we model the density $f(x)$ directly as an SOS polynomial $p(x)$:

$$\begin{aligned}
\text{maximize}_Q \quad & \sum_{i=1}^N \log(v_d(x_i)^T Q v_d(x_i)) \\
\text{subject to} \quad & Q \succeq 0 \\
& \int_\Omega v_d(x)^T Q v_d(x) dx = 1
\end{aligned}$$

The integral constraint is linear in the entries of Q: $\int v_d(x)^T Q v_d(x)dx = \text{Tr}(Q \cdot \int v_d(x)v_d(x)^T dx) = \text{Tr}(Q \cdot M)$, where M is a pre-computable matrix of moments of the Lebesgue measure over Ω. While the objective function here is concave (log-determinant-like), explicitly solving this large-scale SDP can be slow.

3 Applications to "Hard" Problems

Standard density estimation fails when distributions have complex geometries—narrow ridges, high curvature, or multimodality—especially in dimensions $d > 3$.

3.1 Pathological Distributions

In the Bayesian inference and sampling literature, specific distributions are used as benchmarks for difficulty.

3.1.1 The "Banana" Distribution

Often arising in unidentifiable non-linear models, this distribution features a highly curved, non-convex high-density region. Standard Gaussian approximations (Laplace approximation, Variational Inference with mean-field assumptions) fail spectacularly, placing probability mass in the low-density "hole" of the banana or covering only one arm. SOS-based methods can capture this curvature because polynomials naturally model non-linear dependencies between variables (e.g., terms like $x_1^2 x_2^2$ can capture warped correlations).

3.1.2 Neal's Funnel

Introduced by Radford Neal, this is a hierarchical model that wreaks havoc on Markov Chain Monte Carlo (MCMC) samplers:

$$y \sim \mathcal{N}(0, 3^2) \quad \text{and} \quad x_i \mid y \sim \mathcal{N}(0, e^y) \quad \text{for } i = 1, \ldots, d$$

As y becomes negative, the variance of x_i shrinks towards zero, creating an extraordinarily narrow "neck" of probability mass.

- **Why it's hard:** The geometry (curvature) of the distribution changes drastically depending on the region of space. Standard step-size adaptation in samplers fails to handle regions that are simultaneously very broad (top of funnel) and microscopically narrow (bottom of funnel).

- **SOS Solution:** Rather than standard SOS density estimation, modern approaches use SOS for *transport maps* (see below), which can learn to "unwarp" this funnel into a standard Gaussian, effectively learning the complex variable dependencies analytically.

3.2 Sum-of-Squares Polynomial Flows (SOS-PF)

Perhaps the most cutting-edge application is in **Normalizing Flows**. A normalizing flow models a complex density $p_X(x)$ by transforming a simple base density $p_Z(z)$ (e.g., standard Gaussian) through an invertible, differentiable map T:

$$x = T(z) \implies p_X(x) = p_Z(T^{-1}(x)) \left| \det J_{T^{-1}}(x) \right|$$

To be a valid flow, T must be a diffeomorphism. A sufficient condition for easy invertibility is a *triangular* map where each component T_k is monotonic with respect to z_k.

Jaini et al. (2019) proposed *Sum-of-Squares Polynomial Flows*. They ensure $T_k(z_1, \ldots, z_k)$ is strictly increasing in z_k by modeling its partial derivative as an SOS polynomial:

$$\frac{\partial T_k}{\partial z_k} = p_k(z_1, \ldots, z_k) \in \Sigma_{\text{SOS}}$$

Since $p_k \in \Sigma_{\text{SOS}}$, it is guaranteed that $\frac{\partial T_k}{\partial z_k} \geq 0$ everywhere, ensuring monotonicity and invertibility. This allows the model to learn highly complex, multi-modal, and "pathological" target densities by composing several such provably invertible transformations.

4 The Curse of Dimensionality in SOS

While powerful, SOS methods are acutely sensitive to the dimension n of the data. This is not merely a data sparsity issue (classical curse), but a computational complexity issue.

4.1 Combinatorial Explosion of the Basis

The dimension of the vector $v_d(x)$ (and thus the side-length of the Gram matrix Q) is the number of monomials in n variables of degree at most d:

$$N = \dim(\mathbb{R}[x]_{n,d}) = \binom{n+d}{d} = \frac{(n+d)!}{n!d!}$$

Table 1: Growth of SDP problem size with dimension and polynomial degree.

Variables (n)	Degree ($2d$)	Basis size (d)	Matrix Size ($N \times N$)	Approx. Entries
2	4 (quartic)	2	$\binom{4}{2} = 6$	36
5	4 (quartic)	2	$\binom{7}{2} = 21$	441
10	4 (quartic)	2	$\binom{12}{2} = 66$	$\approx 4.3 \times 10^3$
20	6 (sextic)	3	$\binom{23}{3} = 1771$	$\approx 3.1 \times 10^6$

4.2 SDP Solver Complexity

State-of-the-art interior-point methods for generic SDPs have a worst-case time complexity roughly proportional to $O(m^2 N^2 + mN^3)$ or $O(m^{3.5})$ depending on the specific structure, where N is the matrix size and m is the number of constraints. In our case, $N \approx n^d$. The complexity scales roughly as $O(n^{4d})$ or worse. This double-exponential-like growth renders generic SOS density estimation intractable for $n > 15$ or 20 without exploiting specific structural sparsity (e.g., if the variables only interact in small cliques).

Non-negative polynomial density estimation via SOS provides a mathematically rigorous framework for tackling density problems with difficult geometries that defeat simpler local methods. By converting shape constraints into semidefinite constraints, we gain access to global optimization tools. However, the method is inherently limited by the combinatorial growth of the monomial basis, restricting its direct application to low-to-moderate dimensional problems, or requiring modern adaptations like SOS Flows that apply lower-dimensional polynomials iteratively.

5 Density of Sum-of-Squares in Positive Polynomials

A fundamental question in real algebraic geometry and polynomial optimization is whether every non-negative polynomial can be approximated by sum-of-squares (SOS) polynomials. Formally, is the cone of SOS polynomials dense in the cone of positive polynomials?

The answer is **no**, except in very specific trivial cases.

Let $\mathbb{R}[x]_{n,d}$ denote the vector space of real polynomials in n variables of degree d. We identify two convex cones within this space:

- $\mathcal{P}_{n,d}$: The cone of *positive* (non-negative) polynomials, i.e., $\{p \in \mathbb{R}[x]_{n,d} \mid p(x) \geq 0 \quad \forall x \in \mathbb{R}^n\}$.

- $\Sigma_{n,d}$: The cone of *sum-of-squares* polynomials, i.e., $\{s \in \mathbb{R}[x]_{n,d} \mid s(x) = \sum_i q_i(x)^2 \text{ for some } q_i \in \mathbb{R}[x]\}$.

Both $\mathcal{P}_{n,d}$ and $\Sigma_{n,d}$ are closed convex cones. For a subset to be dense in a closed set, it must be equal to that set. Therefore, $\Sigma_{n,d}$ is dense in $\mathcal{P}_{n,d}$ if and only if $\Sigma_{n,d} = \mathcal{P}_{n,d}$.

5.1 Hilbert's 1888 Theorem and Counterexamples

David Hilbert (1888) proved that $\Sigma_{n,d} = \mathcal{P}_{n,d}$ only in three specific cases:

1. $n = 1$: Univariate polynomials of any degree.

2. $d = 2$: Quadratic forms in any number of variables.

3. $n = 2, d = 4$: Bivariate quartic polynomials.

In all other cases, $\Sigma_{n,d} \subsetneq \mathcal{P}_{n,d}$, meaning there exist positive polynomials that are not sums of squares. Because $\Sigma_{n,d}$ is closed, these polynomials cannot even be approximated by SOS polynomials.

The classic counterexample is the **Motzkin polynomial** ($n = 2, d = 6$):

$$M(x,y) = x^4 y^2 + x^2 y^4 - 3x^2 y^2 + 1 \tag{2}$$

By the arithmetic-geometric mean inequality, $M(x,y) \geq 0$ for all $(x,y) \in \mathbb{R}^2$. However, it cannot be written as a sum of squares.

5.2 Guarantees of Positivity on Compact Sets

While SOS is not dense in strictly positive polynomials globally, powerful approximation results—known as *Positivstellensätze*—exist if we restrict the domain to a **compact** semialgebraic set K.

Let $K = \{x \in \mathbb{R}^n \mid g_1(x) \geq 0, \ldots, g_m(x) \geq 0\}$ be a compact set. If a polynomial $f(x)$ is *strictly* positive on K (i.e., $f(x) > 0$ for all $x \in K$), it admits an SOS representation.

Theorem 1 (Schmüdgen's Positivstellensatz, 1991)**.** *If K is compact and $f(x) > 0$ on K, then f can be represented as:*

$$f(x) = \sum_{e \in \{0,1\}^m} \sigma_e(x) \cdot g_1(x)^{e_1} \cdots g_m(x)^{e_m}$$

where each $\sigma_e(x)$ is a sum-of-squares polynomial.

A more practical refinement was provided by Putinar, requiring K to satisfy the slightly stronger *Archimedean condition* (essentially, the constraints g_i must explicitly imply boundedness, e.g., one constraint is $R - \|x\|^2 \geq 0$).

Theorem 2 (Putinar's Positivstellensatz, 1993)**.** *If K is Archimedean and $f(x) > 0$ on K, then f has the simpler linear representation:*

$$f(x) = \sigma_0(x) + \sum_{i=1}^{m} \sigma_i(x) g_i(x)$$

where $\sigma_0, \ldots, \sigma_m$ are sum-of-squares polynomials.

Putinar's theorem is widely used in modern polynomial optimization because searching for the σ_i polynomials can be cast efficiently as a semidefinite program (SDP).

5.3 Global Approximation via Perturbation

To certify non-negativity over all of $\mathbb{R}^n$ (where the Positivstellensätze do not directly apply), one often utilizes perturbation techniques.

If we wish to prove $P(x) \geq 0$ globally, standard SOS relaxations may fail (as with the Motzkin polynomial). A common technique in standard Lasserre hierarchy approaches is to analyze a perturbed polynomial:

$$P_\epsilon(x) = P(x) + \epsilon(1 + \|x\|^2)^k \tag{3}$$

for some small $\epsilon > 0$ and sufficiently large integer k. This perturbation addresses the two main limitations of the classic theorems:

1. **Strict Positivity:** The ϵ term ensures $P_\epsilon(x)$ is strictly positive if $P(x)$ was non-negative.

2. **Coercivity (Compactness substitute):** The term $(1 + \|x\|^2)^k$ grows rapidly as $\|x\| \to \infty$, effectively enforcing an Archimedean-like condition over the unbounded domain $\mathbb{R}^n$.

It has been shown that for strictly positive polynomials, such a perturbed form is guaranteed to be SOS for sufficiently large k, allowing global non-negativity to be certified asymptotically.

References

Jaini, P., Selby, K. A., & Yu, Y. (2019). Sum-of-squares polynomial flow. *International Conference on Machine Learning* (pp. 3009-3018). PMLR.

Lasserre, J. B. (2001). Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization*, 11(3), 796-817.

Neal, R. M. (2003). Slice sampling. *Annals of statistics*, 705-741.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., & Lakshminarayanan, B. (2021). Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57), 1-64.

Parrilo, P. A. (2000). *Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization* (Doctoral dissertation, California Institute of Technology).