Infinite-Dimensional Sums-of-Squares for Optimal Control

Eloïse Berthier

with Ulysse Marteau-Ferey, Justin Carpentier, Alessandro Rudi, Francis Bach

June 17, 2022
Introduction

• We present a representation of non-negative smooth functions in reproducing kernel Hilbert spaces (RKHS), extending the sum-of-squares (SoS) representation of polynomials.
• We apply such representations to optimal control problems, leading to a sample-based numerical method.
Introduction

• We present a representation of non-negative smooth functions in reproducing kernel Hilbert spaces (RKHS), extending the sum-of-squares (SoS) representation of polynomials.
• We apply such representations to optimal control problems, leading to a sample-based numerical method.

The preprint is available on arxiv:
https://arxiv.org/abs/2110.07396
Contents

1 Sums-of-squares Representations for Non-convex Optimization

2 Application to Optimal Control Problems
Non-convex Optimization as a Linear Program

We are interested in finding the global minimum of a possibly non-convex function:

\[ f^* = \min_{x \in \mathbb{R}^p} f(x). \]

This is equivalent to:

\[ f^* = \sup_{c \in \mathbb{R}} c \text{ s.t. } \forall x \in \mathbb{R}^p, f(x) - c \geq 0. \]
Non-convex Optimization as a Linear Program

We are interested in finding the global minimum of a possibly non-convex function:

$$f^* = \min_{x \in \mathbb{R}^p} f(x).$$

This is equivalent to:

$$f^* = \sup_{c \in \mathbb{R}} c \quad \text{s.t.} \quad \forall x \in \mathbb{R}^p, \quad f(x) - c \geq 0.$$
Non-convex Optimization as a Linear Program

We are interested in finding the global minimum of a possibly non-convex function:

\[ f^* = \min_{x \in \mathbb{R}^p} f(x). \]

This is equivalent to:

\[ f^* = \sup_{c \in \mathbb{R}} c \quad \text{s.t.} \quad \forall x \in \mathbb{R}^p, \; f(x) - c \geq 0. \]

How to handle a dense set of constraints?
Idea 1: Subsampling Inequalities

\[ f^* = \sup_{c \in \mathbb{R}} c \]
\[ \text{s.t. } \forall x \in \mathbb{R}^p, f(x) - c \geq 0. \]

Relax it to:

\[ f_n = \sup_{c \in \mathbb{R}} c \]
\[ \text{s.t. } \forall i \in \{1, ..., n\}, f(x_i) - c \geq 0, \]
which is equivalent to...
Idea 1: Subsampling Inequalities

\[ f^* = \sup_{c \in \mathbb{R}} c \]

s.t. \( \forall x \in \mathbb{R}^p, f(x) - c \geq 0 \).

Relax it to:

\[ f_n = \sup_{c \in \mathbb{R}} c \]

s.t. \( \forall i \in \{1, \ldots, n\}, f(x_i) - c \geq 0 \),

which is equivalent to... \( f^* \simeq f_n = \min_i f(x_i) \).
Idea 1: Subsampling Inequalities

\[ f^* = \sup_{c \in \mathbb{R}} c \]
\[ \text{s.t. } \forall x \in \mathbb{R}^p, \ f(x) - c \geq 0. \]

Relax it to:

\[ f_n = \sup_{c \in \mathbb{R}} c \]
\[ \text{s.t. } \forall i \in \{1, \ldots, n\}, \ f(x_i) - c \geq 0, \]

which is equivalent to...

\[ f^* \simeq f_n = \min_i f(x_i). \]

If \( f \) Lipschitz, we need \( O(\varepsilon^{-p}) \) samples to approximate \( f^* \) up to \( \varepsilon \).
If \( f \in C^s(\mathbb{R}^p) \) is smooth, the lower-bound is \( O(\varepsilon^{-p/s}) \) [3].

Can we do any better?
Idea 2: Representing Non-negative Functions

\[ f^* = \sup_{c \in \mathbb{R}} c \]
\[ \text{s.t. } \forall x \in \mathbb{R}^p, \quad f(x) - c \geq 0. \]

We need a practical representation of non-negative functions.
Idea 2: Representing Non-negative Functions

\[ f^* = \sup_{c \in \mathbb{R}} c \]
\[ \text{s.t. } \forall x \in \mathbb{R}^p, f(x) - c \geq 0. \]

We need a \textit{practical representation} of non-negative functions.

Imagine we know how to represent some \( g_k \), \textit{e.g.}, of the form:

\[ g_k(x) = \langle \theta_k, \varphi(x) \rangle. \]

Then we can generate non-negative functions as sum-of-squares:

\[ g(x) = \sum_{k=1}^{m} g_k(x)^2 \]
Idea 2: Representing Non-negative Functions

\[ f^* = \sup_{c \in \mathbb{R}} c \]

\[ \text{s.t. } \forall x \in \mathbb{R}^p, f(x) - c \geq 0. \]

We need a *practical representation* of non-negative functions.

Imagine we know how to represent some \( g_k \), e.g., of the form:

\[ g_k(x) = \langle \theta_k, \varphi(x) \rangle. \]

Then we can generate non-negative functions as sum-of-squares:

\[ g(x) = \sum_{k=1}^{m} g_k(x)^2 = \langle \varphi(x), A \varphi(x) \rangle. \]

where \( A = \sum_{k=1}^{m} \theta_k \otimes \theta_k \geq 0 \) has rank less than \( m \).
Polynomial Sum-of-Squares (SoS)

In dimension 1, all non-negative polynomials are SoS. This is not true in larger dimensions.

Powerful theorems describe cases of tight SoS representations [1].
Polynomial Sum-of-Squares (SoS)

In dimension 1, all non-negative polynomials are SoS. This is not true in larger dimensions.

Powerful theorems describe cases of tight SoS representations [1].

**Theorem (Putinar’s Positivstellensatz (simplified))**

Let \((h_k)_k\) a family of polynomials and \(W = \{x \mid \forall k, h_k(x) \geq 0\}\) a semi-algebraic set. Assume that \(\{x \in \mathbb{R}^d \mid h_k(x) \geq 0\}\) is compact for some \(k\). If a polynomial \(f\) is strictly positive on \(W\), then there exists SoS polynomials \((\sigma_k)_{0 \leq k \leq m}\) such that:

\[
f = \sigma_0 + \sum_{k=1}^{m} \sigma_k h_k.
\]
The Moment – SoS Hierarchy

Let \( f \) a polynomial of degree \( d_0 \). We want to solve:

\[
\min_{x \in \mathbb{R}^p} f(x) \quad \text{s.t.} \quad \forall k \in \{1, \ldots, m\}, \quad h_k(x) \geq 0.
\]

Lasserre’s Hierarchy of semi-definite programs (SDP):

Find \( c, \ X_k \succeq 0, \ k = 0, \ldots, m \) such that

\[
\forall \alpha \in \mathbb{N}^p_{d_0}, \quad f_\alpha - c \mathbf{1}_{\alpha=0} = \sum_{k=0}^{m} \langle C^k_\alpha, X_k \rangle
\]

\[
\forall \alpha \in \mathbb{N}^p_{2r} \setminus \mathbb{N}^p_{d_0}, \quad 0 = \sum_{k=0}^{m} \langle C^k_\alpha, X_k \rangle.
\]
The Moment – SoS Hierarchy

Let \( f \) a polynomial of degree \( d_0 \). We want to solve:

\[
f^* = \min_{x \in \mathbb{R}^p} f(x) \text{ s.t. } \forall k \in \{1, ..., m\}, \ h_k(x) \geq 0.
\]

Lasserre’s Hierarchy of semi-definite programs (SDP):

Find \( c, X_k \succeq 0, k = 0, ..., m \) such that

\[
\forall \alpha \in \mathbb{N}^p_{d_0}, \quad f_\alpha - c1_{\alpha=0} = \sum_{k=0}^m \langle C^k_\alpha, X_k \rangle
\]

\[
\forall \alpha \in \mathbb{N}^p_{2r} \setminus \mathbb{N}^p_{d_0}, \quad 0 = \sum_{k=0}^m \langle C^k_\alpha, X_k \rangle.
\]

The monomials are indexed by \( \mathbb{N}^p_r := \{ \alpha \in \mathbb{N}^p : |\alpha| \leq r \} \) which has size \( s_r(d) = \binom{p+r}{r} \). This is exponential in the dimension \( p \).
SoS Functions in Reproducing Kernel Hilbert Spaces

A positive-definite kernel on $\mathbb{R}^p$ is a function $K : \mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R}$ such that $\forall n \geq 1, \forall (x_1, ..., x_n)$, the matrix $(K(x_i, x_j))$ is PSD.

It is associated to a Hilbert space $\mathcal{H}$ such that:

- $\forall x \in \mathbb{R}^p$, $\varphi(x) := K(x, \cdot) \in \mathcal{H}$;
- $\forall f \in \mathcal{H}, x \in \mathbb{R}^p$, $\langle f, \varphi(x) \rangle = f(x)$. 

---

Eloïse Berthier, with Ulysse Marteau-Ferey, Justin Carpentier, Alessandro Rudi, Francis Bach - Infinite-Dimensional Sums-of-Squares for Optimal Control
SoS Functions in Reproducing Kernel Hilbert Spaces

A positive-definite kernel on $\mathbb{R}^p$ is a function $K : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ such that $\forall n \geq 1, \forall (x_1, ..., x_n)$, the matrix $(K(x_i, x_j))$ is PSD.

It is associated to a Hilbert space $\mathcal{H}$ such that:

- $\forall x \in \mathbb{R}^p$, $\varphi(x) := K(x, \cdot) \in \mathcal{H}$;
- $\forall f \in \mathcal{H}, x \in \mathbb{R}^p$, $\langle f, \varphi(x) \rangle = f(x)$.

Using the reproducing property, a sum-of-squares of functions in $\mathcal{H}$:

$$\forall x \in \mathbb{R}^p, \quad g(x) = \sum_{k=1}^{m} h_k(x)^2$$

is such that

$$\forall x \in \mathbb{R}^d, \quad g(x) = \langle \varphi(x), A\varphi(x) \rangle,$$

where $A \in \mathbb{S}_+(\mathcal{H})$ is a PSD operator, possibly infinite-dimensional.
SoS Representation of Smooth Functions [4]

We consider as our RKHS $\mathcal{H}$ the Sobolev space $W^s_2(\mathbb{R}^p)$, with $s \geq p/2 + 3$, of $s$-smooth functions.

**Theorem (informal)**

If $f \in \mathcal{H}$ has a unique isolated global minimum at $x^*$ s.t. $\frac{\partial^2 f}{\partial x^2}(x^*) \succ 0$, then there exists $h_1, \ldots, h_m \in \mathcal{H}$, with $m \leq p + 1$ such that:

$$\forall x, \quad f(x) - f^* = \sum_{k=1}^{m} h_k(x)^2.$$ 

Hence $f - f^*$ is a SoS of (smooth) functions in $\mathcal{H}$, and:

$$\exists A \in \mathbb{S}_+(\mathcal{H}) \text{ s.t. } \forall x, \quad f(x) - f^* = \langle \varphi(x), A\varphi(x) \rangle.$$ 

No need for a hierarchy!
Non-convex Optimization of Smooth Functions

\[ f^* = \sup_{c \in \mathbb{R}} c \]

s.t. \( \forall x \in \mathbb{R}^p, f(x) - c \geq 0. \)

Using the Theorem, if \( f \) is smooth, this is equivalent to:

\[ f^* = \sup_{c \in \mathbb{R}, A \in \mathcal{S}_+(\mathcal{H})} c \]

s.t. \( \forall x, f(x) - c = \langle \varphi(x), A \varphi(x) \rangle. \)
Non-convex Optimization of Smooth Functions

We can now subsample equalities:

\[
f_n = \sup_{c \in \mathbb{R}, A \in \mathcal{S}_+(\mathcal{H})} c - \lambda \text{Tr}(A)
\]

s.t. \( \forall i \in \{1, \ldots, n\}, f(x_i) - c = \langle \varphi(x_i), A\varphi(x_i) \rangle \).

Using the reproducing property, this is equivalent to the SDP:

\[
f_n = \sup_{c \in \mathbb{R}, B \succeq 0} c - \lambda \text{Tr}(B)
\]

s.t. \( \forall i \in \{1, \ldots, n\}, f(x_i) - c = \Phi^\top_i B\Phi_i \),

where the \( \Phi_i \in \mathbb{R}^n \) are vectors computed from the kernel matrix.

This achieves an almost optimal rate of \( O\left(\frac{n}{s - 3} + \frac{1}{2}\right) \) for \( s \geq 3 + \frac{p}{2} \). The lower-bound is \( O\left(\frac{n}{s/p}\right) \).
Non-convex Optimization of Smooth Functions

We can now *subsample equalities*:

\[
  f_n = \sup_{c \in \mathbb{R}, A \in \mathbb{S}_+^n(\mathcal{H})} c - \lambda \text{Tr}(A)
\]

s.t. \( \forall i \in \{1, \ldots, n\}, \; f(x_i) - c = \langle \varphi(x_i), A\varphi(x_i) \rangle \).

Using the reproducing property, this is equivalent to the SDP:

\[
  f_n = \sup_{c \in \mathbb{R}, B \succeq 0} c - \lambda \text{Tr}(B)
\]

s.t. \( \forall i \in \{1, \ldots, n\}, \; f(x_i) - c = \Phi_i^\top B \Phi_i \),

where the \( \Phi_i \in \mathbb{R}^n \) are vectors computed from the kernel matrix.

This achieves an *almost optimal rate* of \( O(n^{-(s-3)/p+1/2}) \) for \( s \geq 3 + p/2 \). The lower-bound is \( O(n^{-s/p}) \).
Contents

1 Sums-of-squares Representations for Non-convex Optimization

2 Application to Optimal Control Problems
Optimal Control as a Linear Program

- The optimal control problem is to find $V^*$ such that:

$$V^*(t_0, x_0) = \inf_{u(\cdot)} \int_{t_0}^{T} L(t, x(t), u(t)) dt + M(x(T))$$

$$\forall t \in [t_0, T], \quad \dot{x}(t) = f(t, x(t), u(t)), \quad x(0) = x_0.$$
The optimal control problem is to find $V^*$ such that:

$$V^*(t_0, x_0) = \inf_{u(\cdot)} \int_{t_0}^{T} L(t, x(t), u(t)) dt + M(x(T))$$

$$\forall t \in [t_0, T], \dot{x}(t) = f(t, x(t), u(t)), \quad x(0) = x_0.$$  

Under convexity assumptions, this is equivalent to finding a maximal subsolution of the Hamilton-Jacobi-Bellman equation [2]:

$$\sup_{V \in C^1([0, T] \times \mathcal{X})} \int V(0, x_0) d\mu_0(x_0)$$

$$\forall (t, x, u), \frac{\partial V}{\partial t}(t, x) + L(t, x, u) + \nabla V(t, x)^\top f(t, x, u) \geq 0$$

$$\forall x, V(T, x) \leq M(x).$$
Optimal Control as a Linear Program

• The optimal control problem is to find $V^*$ such that:

$$V^*(t_0, x_0) = \inf_{u(\cdot)} \int_{t_0}^{T} L(t, x(t), u(t)) dt + M(x(T))$$

$$\forall t \in [t_0, T], \dot{x}(t) = f(t, x(t), u(t)), \quad x(0) = x_0.$$ 

• Under convexity assumptions, this is equivalent to finding a *maximal subsolution* of the Hamilton-Jacobi-Bellman equation [2]:

$$\sup_{V \in C^1([0, T] \times X)} \int V(0, x_0) d\mu_0(x_0)$$

$$\forall (t, x, u), \left(\frac{\partial V}{\partial t} (t, x) + L(t, x, u) + \nabla V(t, x)^\top f(t, x, u) \geq 0 \right)$$

$$\forall x, V(T, x) \leq M(x).$$
A Simple Baseline: Subsampling Inequalities

Using a linear parameterization of $V$, and simply subsampling inequalities leads to an LP:

$$\sup_{\theta \in \mathbb{R}^m} \frac{1}{n} \sum_{i=1}^{n} V_{\theta}(0, x^{(i)})$$

$$\forall i \in I, \quad H_{\theta}(t^{(i)}, x^{(i)}, u^{(i)}) \geq 0.$$  

This is already a non-trivial numerical method.

*Can we do any better?*
SoS Representation of the Hamiltonian

**Theorem (informal)**

Assume that:

- \( f \) is **control-affine**: 
  \[ f(t, x, u) = g(t, x) + B(t, x)u; \]
- \( L \) is strongly convex in \( u \);
- \( L, B \) and \( V^* \) are **sufficiently smooth**;

Then \( H^* \) is a SoS of \( p \) smooth functions \( (w_j)_{1 \leq j \leq p} \in C^s(\Omega) \):

\[
\forall (t, x, u) \in \Omega, \quad H^*(t, x, u) = \sum_{j=1}^{p} w_j(t, x, u)^2.
\]

**Limit**: in general \( V^* \) is not even \( C^1 \). A possible workaround is to add noise to the dynamics to smoothen \( V^* \).
A Practical Algorithm for Smooth Optimal Control

• Adding regularization terms, we get the following SDP:

\[
\sup_{B \succeq 0, \theta, \delta} \quad c^\top \theta - \lambda_{\theta} \| \theta \|^2_2 - \lambda \text{Tr}(B) - \gamma \| \delta \|^2 + \varepsilon \log \det B + C
\]

such that \( \forall i \in \{1, \ldots, n\}, \ b_i + a_i^\top \theta = (\Phi_i)^\top B \Phi_i + \delta_i. \)

• The dual is solved with damped Newton’s method, an algorithm with cost \( O(n^3) \) in time and \( O(n^2) \) in space for each iteration.
Numerical Example

We solve a linear quadratic regulator. We know that:

$$H^*(t, x, u) = (u + F(t)x)^\top R(u + F(t)x).$$

We use:

$$K((t, x, u), (t', x', u')) = \langle u, u' \rangle + \langle x, x' \rangle e^{-|t-t'|}.$$
Numerical Example

We solve a linear quadratic regulator. We know that:

\[ H^*(t, x, u) = (u + F(t)x)^\top R(u + F(t)x). \]

We use:

\[ K((t, x, u), (t', x', u')) = \langle u, u' \rangle + \langle x, x' \rangle e^{-|t-t'|}. \]
Conclusion

- We have presented an extension of the SoS framework in RKHS.
- It leads to sample-based numerical methods involving SDPs.
- There are many potential applications, e.g., to optimal transport, sampling, modelling of probability distributions...
References

J.-B. Lasserre.


E. Novak.


A. Rudi, U. Marteau-Ferey, and F. Bach.