Polynomial optimization on finite sets.

Mauricio Velasco
Universidad Católica del Uruguay (UCU)

RECO2023
Pontificia Universidad Católica

LECTURE 4: The polynomial kernel method.

Let n be an even integer and $X := \{-1, 1\}^n$. We know that for every quadratic function $f \in \mathbb{R}[X]_{\leq 2}$ and $r \geq \frac{n}{2} + 1$ the equality $f_{\min} = f_{(r)}$ holds where $f_{(r)}$ is the semidefinite programming lower bound of level r,

$$f_{(r)} := \max\{\lambda : f - \lambda \in \Sigma_{\leq r}\}$$

Let n be an even integer and $X := \{-1,1\}^n$. We know that for every quadratic function $f \in \mathbb{R}[X]_{\leq 2}$ and $r \geq \frac{n}{2}+1$ the equality $f_{\min} = f_{(r)}$ holds where $f_{(r)}$ is the semidefinite programming lower bound of level r,

$$f_{(r)} := \max\{\lambda : f - \lambda \in \Sigma_{\leq r}\}$$

Can we bound the gap $f_{\min} - f_{(r)}$ as a function of r?

Let n be an even integer and $X := \{-1, 1\}^n$. We know that for every quadratic function $f \in \mathbb{R}[X]_{\leq 2}$ and $r \geq \frac{n}{2} + 1$ the equality $f_{\min} = f_{(r)}$ holds where $f_{(r)}$ is the semidefinite programming lower bound of level r,

$$f_{(r)} := \max \{\lambda : f - \lambda \in \Sigma_{\leq r}\}$$

Can we bound the gap $f_{\min} - f_{(r)}$ as a function of r?
More precisely we would like to bound the worst-case gap

$$\sup_{f \in \mathbb{R}[X]_{\leq 2}} \frac{f_{\min} - f_{(r)}}{\|f\|_\infty} \leq F(r)$$

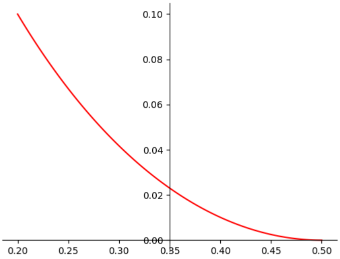
There is a good answer to this question for every d, There is a good answer to this question for every d,

Theorem. (Laurent, Slot (2021))

Suppose $f \in \mathbb{R}[X]_{\leq d}$. There exists a constant $C(d)$ such that

$$\frac{f_{\min} - f_{(r)}}{\|f\|_\infty} \leq C(d)\frac{\zeta_r}{n}$$

where ζ_r is the smallest root of the Krawtchouk polynomial $K_r(t)$. Furthermore $\frac{\zeta_r}{n} \sim \phi(r/n)$.



Theorem. (Laurent, Slot (2021))

Suppose $f \in \mathbb{R}[X]_{\leq d}$. There exists a constant $C(d)$ such that

$$\frac{f_{\min} - f_{(r)}}{\|f\|_\infty} \leq C(d)\frac{\zeta_r}{n}$$

where ζ_r is the smallest root of the Krawtchouk polynomial $K_r(t)$

- The main idea of the proof is to **perturb** nonnegative polynomials until they become sums-of-squares of low degree and to estimate the size of this perturbation.

Theorem. (Laurent, Slot (2021))

Suppose $f \in \mathbb{R}[X]_{\leq d}$. There exists a constant $C(d)$ such that

$$\frac{f_{\min} - f_{(r)}}{\|f\|_\infty} \leq C(d)\frac{\zeta_r}{n}$$

where ζ_r is the smallest root of the Krawtchouk polynomial $K_r(t)$

- The main idea of the proof is to **perturb** nonnegative polynomials until they become sums-of-squares of low degree and to estimate the size of this perturbation.
- Such perturbations are built by constructing **local averages** of functions, via polynomial kernels.

Theorem. (Laurent, Slot (2021))

Suppose $f \in \mathbb{R}[X]_{\leq d}$. There exists a constant $C(d)$ such that

$$\frac{f_{\min} - f_{(r)}}{\|f\|_\infty} \leq C(d)\frac{\zeta_r}{n}$$

where ζ_r is the smallest root of the Krawtchouk polynomial $K_r(t)$

- The main idea of the proof is to **perturb** nonnegative polynomials until they become sums-of-squares of low degree and to estimate the size of this perturbation.
- Such perturbations are built by constructing **local averages** of functions, via polynomial kernels.
- There are many possible polynomial kernels and a **good choice**, taking advantage of the symmetries of the problem leads to the proof of the Theorem.

Plan for Lecture 4:

1. The polynomial kernel method.
2. Invariant polynomial kernels on the hypercube.
3. Generalizing the hypercube.

Part 1:
Local averaging via polynomial kernels.

Suppose $X \subseteq \mathbb{R}^n$ is a metric space having distance function $d(x, y)$ and a given probability measure μ.

Suppose $X \subseteq \mathbb{R}^n$ is a metric space having distance function $d(x, y)$ and a given probability measure μ.

A polynomial on X can be a very complicated function, so we will pass it through a polynomial *low pass filter*...

Suppose $X \subseteq \mathbb{R}^n$ is a metric space having distance function $d(x, y)$ and a given probability measure μ.

A polynomial on X can be a very complicated function, so we will pass it through a polynomial *low pass filter*...

This filter is built using the distance function and an auxiliary polynomial $g(t)$.

Suppose $X \subseteq \mathbb{R}^n$ is a metric space having distance function $d(x, y)$ and a given probability measure μ.

A polynomial on X can be a very complicated function, so we will pass it through a polynomial *low pass filter*...

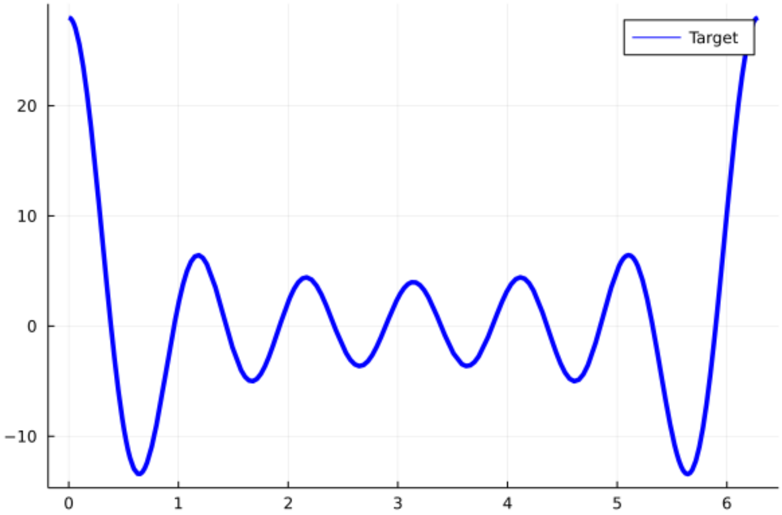
This filter is built using the distance function and an auxiliary polynomial $g(t)$.

Definition.

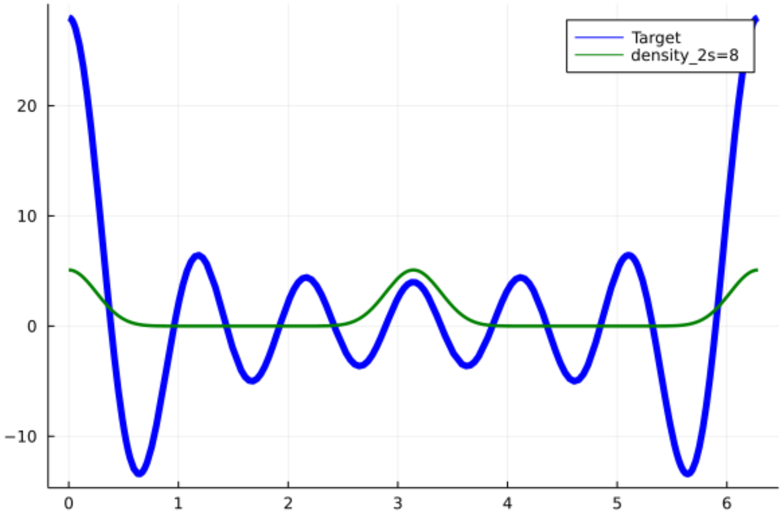
Let $g(t)$ be a univariate sum-of-squares. Define $\Gamma_g : \mathbb{R}[X] \to \mathbb{R}[X]$ via $\Gamma_g(f(x)) = h(x)$ where

$$h(x) = \int_X g\left(d(x, y)\right) f(y) d\mu(y).$$

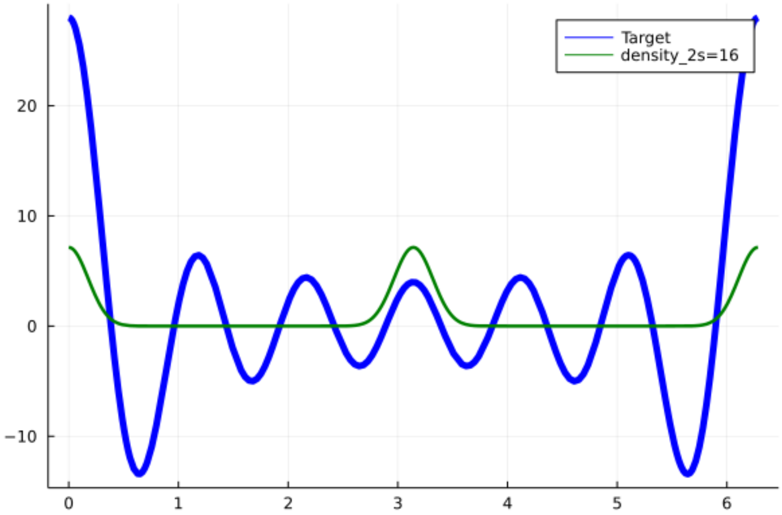
Averaging polynomials



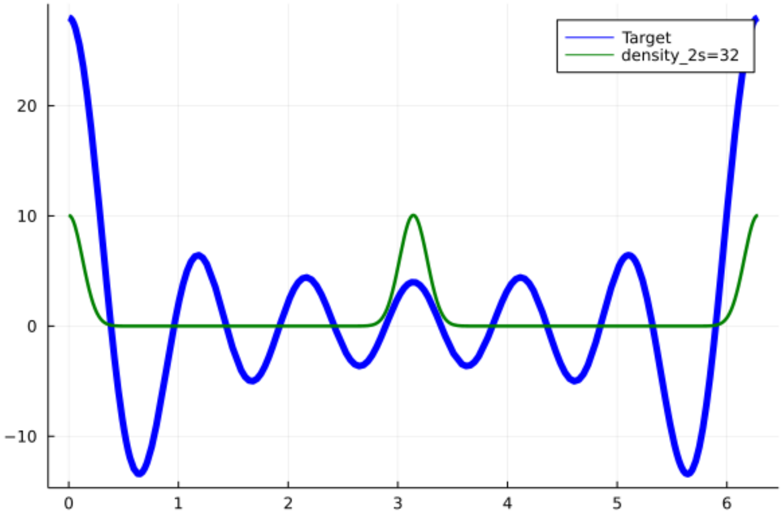
Averaging polynomials



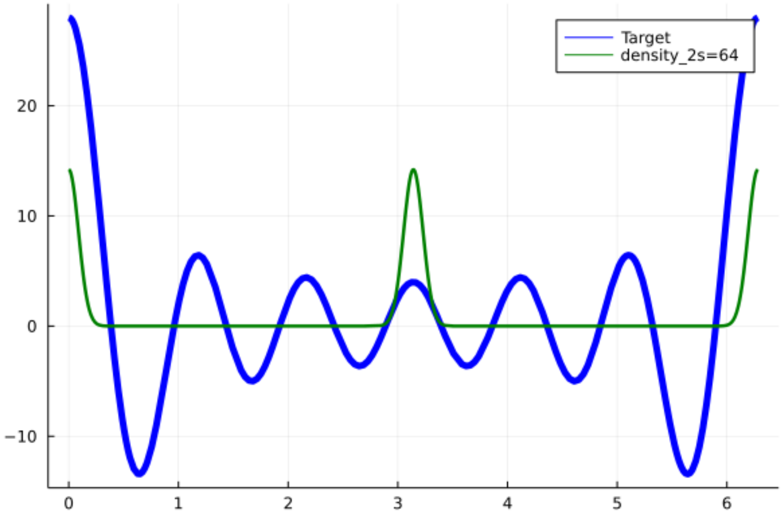
Averaging polynomials



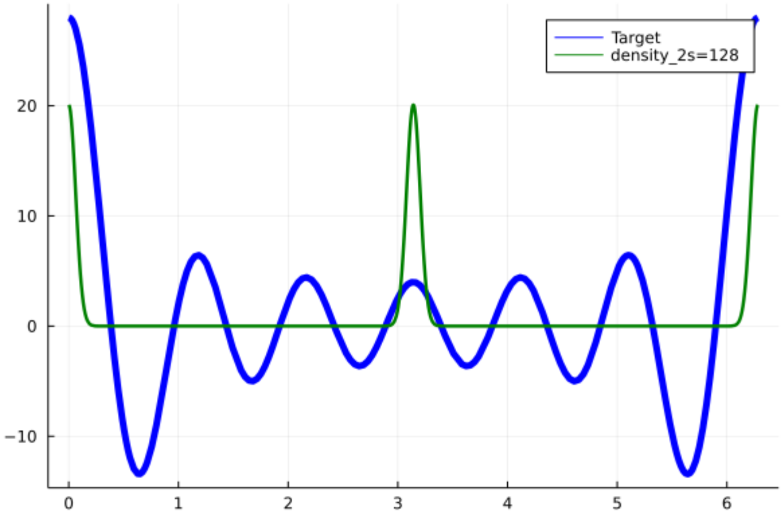
Averaging polynomials



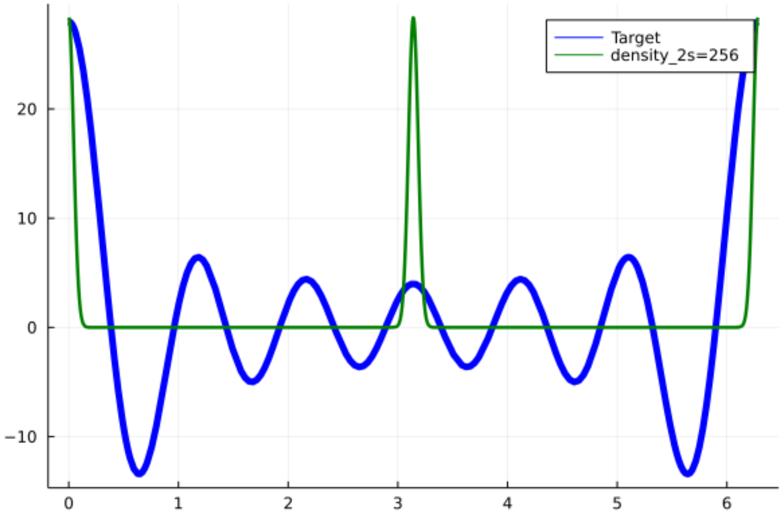
Averaging polynomials



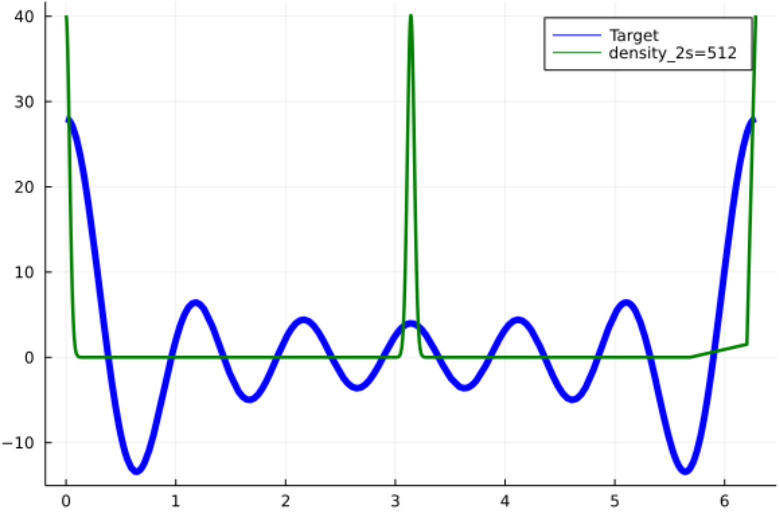
Averaging polynomials



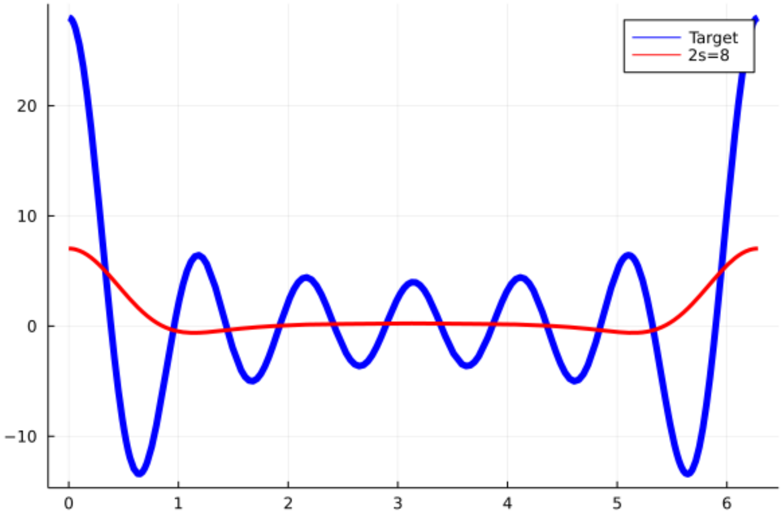
Averaging polynomials



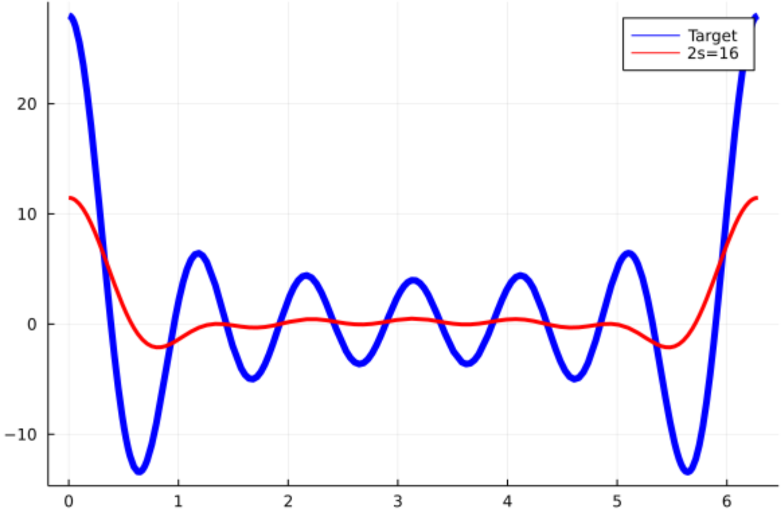
Averaging polynomials



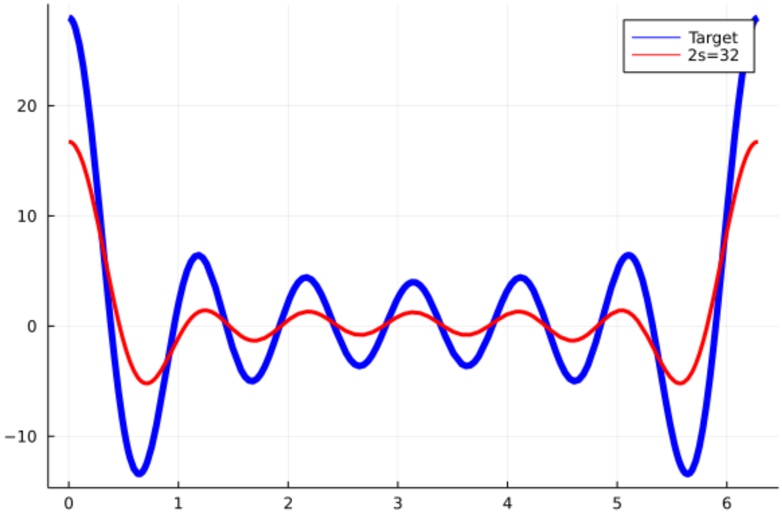
Averaging polynomials



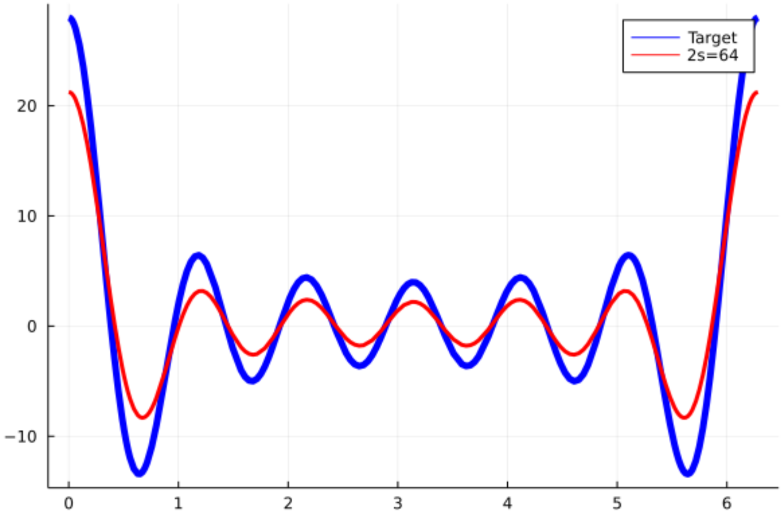
Averaging polynomials



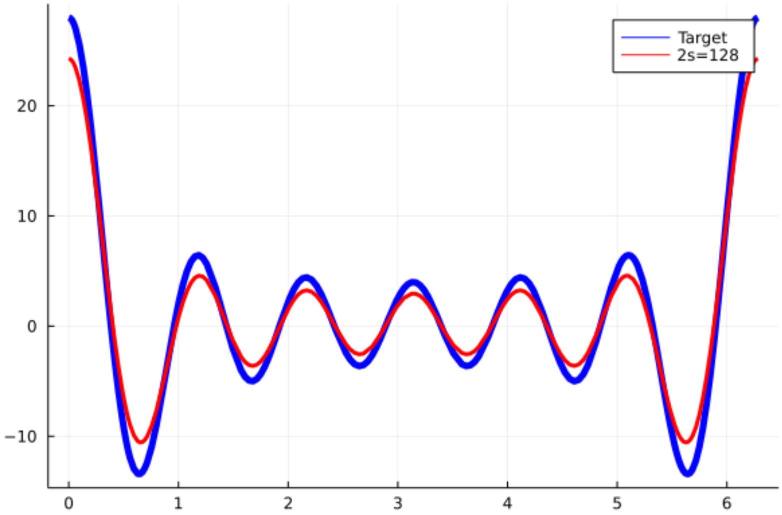
Averaging polynomials



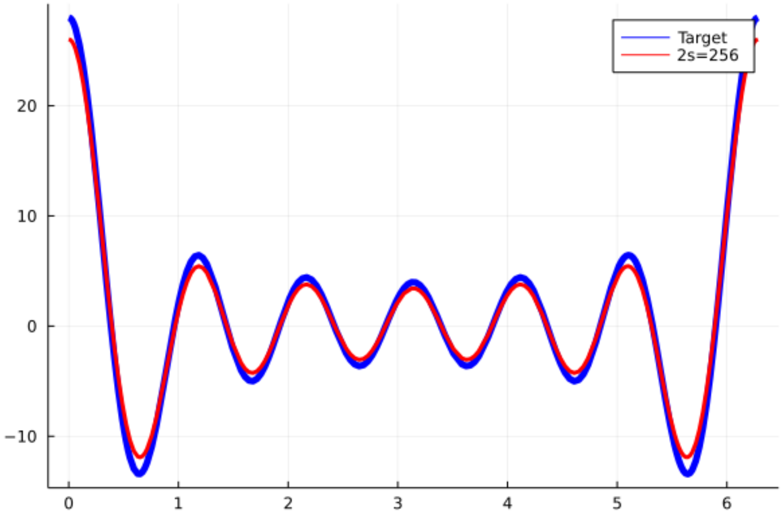
Averaging polynomials



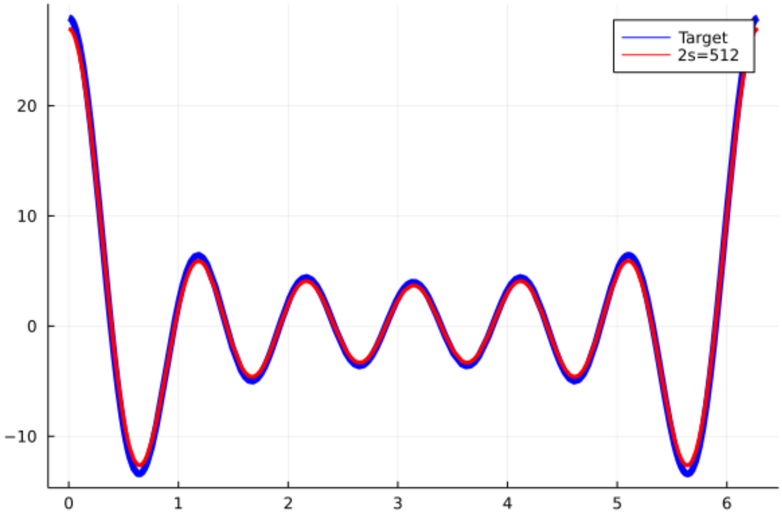
Averaging polynomials



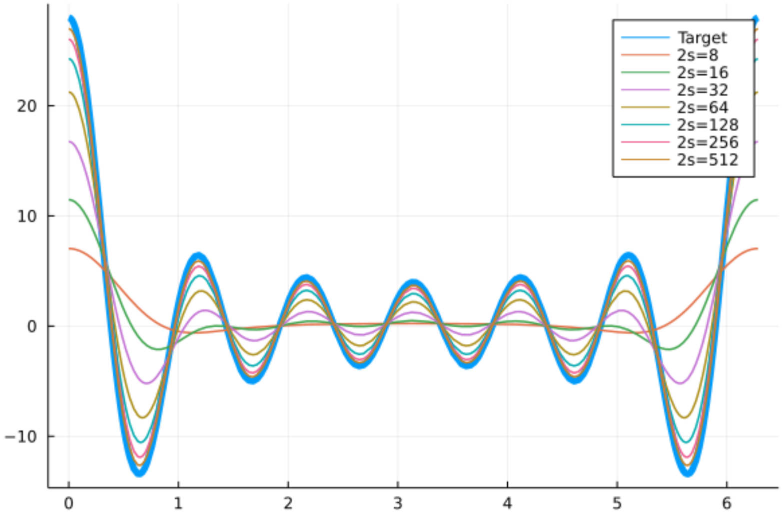
Averaging polynomials



Averaging polynomials



Averaging polynomials



Why are averaging operations useful?

After $f(x)$ *goes through the filter, the resulting polynomial* $h(x)$ *typically becomes simpler...*

Why are averaging operations useful?

After $f(x)$ goes through the filter, the resulting polynomial $h(x)$ typically becomes simpler...

Assume X is finite, μ is the counting measure and for every y_0 the function $d(x, y_0)$ is affine linear.

Lemma.

Assume $g(t) = s(t)^2$ is a square of a polynomial of degree $\leq r$. If $f(x)$ is nonnegative on X then $h(x)$ is a sum-of-squares of functions of degree $\leq r$.

Why are averaging operations useful?

After $f(x)$ goes through the filter, the resulting polynomial $h(x)$ typically becomes simpler...

Assume X is finite, μ is the counting measure and for every y_0 the function $d(x, y_0)$ is affine linear.

Lemma.

Assume $g(t) = s(t)^2$ is a square of a polynomial of degree $\leq r$. If $f(x)$ is nonnegative on X then $h(x)$ is a sum-of-squares of functions of degree $\leq r$.

Proof.

$$h(x) = \int_X s_i(d(x, y))^2 f(y) d\mu(y) = \frac{1}{|X|} \sum_{y_0 \in X} s_i(d(x, y_0))^2 f(y_0)$$

$\square$

Why are averaging operations useful?

If there exists a filter Γ_g *with* $g(t) = s(t)^2$ *which is* **close to the identity** *operator then small perturbations of nonnegative polynomials are sums of squares*

Why are averaging operations useful?

If there exists a filter Γ_g with $g(t) = s(t)^2$ which is **close to the identity** *operator then small perturbations of nonnegative polynomials are sums of squares*

To measure distances between operators $L : \mathbb{R}[X] \to \mathbb{R}[X]$ we will use the operator norm

$$\|L\| := \sup_{\|p\|_\infty \leq 1} \|L(p)\|_\infty$$

where $\|p\|_\infty := \sup_{x \in X} |p(x)|$.

Theorem. (Reznick / Fang-Fawzi)

Assume $g(t) = s(t)^2$ is the square of a polynomial of degree $\leq r$. If $\Gamma_g(1) = 1$ and $\|\Gamma_g^{-1} - I\| \leq \delta$ then $\sup_f \frac{f_{\min} - f_{(r)}}{\|f\|_\infty} \leq \delta$.

Theorem. (Reznick / Fang-Fawzi)

Assume $g(t) = s(t)^2$ *is the square of a polynomial of degree* $\leq r$.
If $\Gamma_g(1) = 1$ *and* $\|\Gamma_g^{-1} - I\| \leq \delta$ *then* $\sup_f \frac{f_{\min} - f_{(r)}}{\|f\|_\infty} \leq \delta$.

Proof.

We will prove that $f - f_{\min} + \delta \in \Sigma_{\leq r}$ if $\|f\|_\infty \leq 1$.

Theorem. (Reznick / Fang-Fawzi)

Assume $g(t) = s(t)^2$ *is the square of a polynomial of degree* $\leq r$*.*
If $\Gamma_g(1) = 1$ *and* $\|\Gamma_g^{-1} - I\| \leq \delta$ *then* $\sup_f \frac{f_{\min} - f_{(r)}}{\|f\|_\infty} \leq \delta$*.*

Proof.

We will prove that $f - f_{\min} + \delta \in \Sigma_{\leq r}$ if $\|f\|_\infty \leq 1$.

$$\Gamma_g^{-1}(f - f_{\min} + \delta) = \Gamma_g^{-1}(f) - f_{\min} + \delta =$$

Theorem. (Reznick / Fang-Fawzi)

Assume $g(t) = s(t)^2$ *is the square of a polynomial of degree* $\leq r$*.*
If $\Gamma_g(1) = 1$ *and* $\|\Gamma_g^{-1} - I\| \leq \delta$ *then* $\sup_f \frac{f_{\min} - f_{(r)}}{\|f\|_\infty} \leq \delta$*.*

Proof.

We will prove that $f - f_{\min} + \delta \in \Sigma_{\leq r}$ if $\|f\|_\infty \leq 1$.

$$\Gamma_g^{-1}(f - f_{\min} + \delta) = \Gamma_g^{-1}(f) - f_{\min} + \delta =$$

$$= \Gamma_g^{-1}(f) - f_{\min} + \delta \geq f - \delta - f_{\min} + \delta = f - f_{\min} \geq 0$$

The result follows since the nonnegativity of $\Gamma_g^{-1}(h)$ implies that $h = \Gamma_g(\Gamma_g^{-1}(h)) \in \Sigma_{\leq r}$. □

Let's recapitulate...

Let's recapitulate...

If $g(t) = s^2(t)$ *with* $\deg(s) \leq r$ *satisfies* $\Gamma_g(1) = 1$ *then we have the inequality*

$$\frac{f_{\min} - f_{(r)}}{\|f\|_\infty} \leq \|\Gamma_g^{-1} - I\|$$

Let's recapitulate...

If $g(t) = s^2(t)$ *with* $\deg(s) \leq r$ *satisfies* $\Gamma_g(1) = 1$ *then we have the inequality*

$$\frac{f_{\min} - f_{(r)}}{\|f\|_\infty} \leq \|\Gamma_g^{-1} - I\|$$

This suggests the following two questions:

1. Given $g(t)$, how to compute $\|\Gamma_g^{-1} - I\|$?
2. How to choose $g(t)$ so $\|\Gamma_g^{-1} - I\|$ is small?

Let's recapitulate...

If $g(t) = s^2(t)$ with $\deg(s) \leq r$ satisfies $\Gamma_g(1) = 1$ then we have the inequality

$$\frac{f_{\min} - f_{(r)}}{\|f\|_\infty} \leq \|\Gamma_g^{-1} - I\|$$

This suggests the following two questions:

1. Given $g(t)$, how to compute $\|\Gamma_g^{-1} - I\|$?
2. How to choose $g(t)$ so $\|\Gamma_g^{-1} - I\|$ is small?

For the hypercube $X = \{-1, 1\}^n \subseteq \mathbb{R}^n$ both questions have good answers...

Part 2:
Invariant kernels on the hypercube $X := \{-1, 1\}^n$.

Theorem. (Laurent, Slot, 2021)

There exists a collection of univariate orthogonal polynomials $\hat{K}_j(t)$ for $j = 0, \ldots, n$ and a decomposition

$$\mathbb{R}[X] = W_0 \oplus \cdots \oplus W_n$$

into orthogonal subspaces having the following property: If $g(t) = \sum \lambda_i \hat{K}_i(t)$ is the unique expression of a polynomial $g(t)$ then, in any basis for $\mathbb{R}[X]$ compatible with the above decomposition we have

$$[\Gamma_g] = \begin{pmatrix} \lambda_1 I_{d_1} & 0 & 0 & \ldots & 0 \\ 0 & \lambda_2 I_{d_2} & 0 & \ldots & 0 \\ 0 & 0 & \lambda_3 I_{d_3} & \ldots & 0 \\ \vdots & \vdots & \vdots & \ldots & 0 \\ 0 & 0 & \ldots & \ldots & \lambda_n I_{d_n} \end{pmatrix}$$

Why is this useful?

- The block-diagonal structure of Γ_g allows us to get a simple estimate of the norm [Fang-Fawzi],

$$\|\Gamma_g^{-1} - I\| \leq \gamma \sum_{i=1}^{n} (1 - \lambda_i)$$

Why is this useful?

- The block-diagonal structure of Γ_g allows us to get a simple estimate of the norm [Fang-Fawzi],

$$\|\Gamma_g^{-1} - I\| \leq \gamma \sum_{i=1}^{n} (1 - \lambda_i)$$

- This estimate is **linear** in the λ_i allowing us to formulate an optimization problem (SDP) which **searches** for good $g(t)$

Why is this useful?

- The block-diagonal structure of Γ_g allows us to get a simple estimate of the norm [Fang-Fawzi],

$$\|\Gamma_g^{-1} - I\| \leq \gamma \sum_{i=1}^{n} (1 - \lambda_i)$$

- This estimate is **linear** in the λ_i allowing us to formulate an optimization problem (SDP) which **searches** for good $g(t)$

$$\min_{g(t)} \left\{ \sum_{j=1}^{n} \left(1 - \langle \hat{K}_j(t), g(t) \rangle\right) : g(t) \in \Sigma_{\leq r}^{\mathbb{R}[t]} \ , \ \langle 1, g(t) \rangle = 1 \right\}$$

Why is this useful?

- The block-diagonal structure of Γ_g allows us to get a simple estimate of the norm [Fang-Fawzi],

$$\|\Gamma_g^{-1} - I\| \leq \gamma \sum_{i=1}^{n} (1 - \lambda_i)$$

- This estimate is **linear** in the λ_i allowing us to formulate an optimization problem (SDP) which **searches** for good $g(t)$

$$\min_{g(t)} \left\{ \sum_{j=1}^{n} \left(1 - \langle \hat{K}_j(t), g(t) \rangle\right) : g(t) \in \Sigma_{\leq r}^{\mathbb{R}[t]} \ , \ \langle 1, g(t) \rangle = 1 \right\}$$

With the inner product that makes the $\hat{K}_j$ orthonormal.

- The optimization problem

$$\min_{g(t)} \left\{ \sum_{j=1}^{n} \Big(1 - \langle \hat{K}_j(t), g(t) \rangle \Big) : g(t) \in \Sigma_{\leq r}^{\mathbb{R}[t]} \,,\; \langle 1, g(t) \rangle = 1 \right\}$$

is a **univariate sum-of-squares problem** (SDP) involving matrices of size $(r+1) \times (r+1)$ with $r \leq n/2$.

- The optimization problem

$$\min_{g(t)} \left\{ \sum_{j=1}^{n} \left(1 - \langle \hat{K}_j(t), g(t) \rangle\right) : g(t) \in \Sigma_{\leq r}^{\mathbb{R}[t]} \ , \ \langle 1, g(t) \rangle = 1 \right\}$$

  is a **univariate sum-of-squares problem** (SDP) involving matrices of size $(r+1) \times (r+1)$ with $r \leq n/2$.
- The Theorem of Laurent-Slot results from an explicit analysis of the SDP via the combinatorics of Krawtchouk polynomials.

- The optimization problem

$$\min_{g(t)} \left\{ \sum_{j=1}^{n} \left(1 - \langle \hat{K}_j(t), g(t) \rangle\right) : g(t) \in \Sigma_{\leq r}^{\mathbb{R}[t]} \ , \ \langle 1, g(t) \rangle = 1 \right\}$$

  is a **univariate sum-of-squares problem** (SDP) involving matrices of size $(r+1) \times (r+1)$ with $r \leq n/2$.
- The Theorem of Laurent-Slot results from an explicit analysis of the SDP via the combinatorics of Krawtchouk polynomials.

Remark.

The analysis of Laurent-Slot is a discrete analogue to the work of [Fang-Fawzi, 2020] on the sphere.

Remark.

Explicit polynomial kernels combined with quadrature rules can be used to create novel optimization algorithms on spaces admitting both (see [Cristancho, -, 2022] on the sphere).

Why do Krawtchouk polynomials exist?

- We think of the hypercube $X = \{-1, 1\}^n$ as a metric space with the Hamming distance:

$$d(x, y) = \#\{i \in [n] : x_i \neq y_i\}$$

Why do Krawtchouk polynomials exist?

- We think of the hypercube $X = \{-1, 1\}^n$ as a metric space with the Hamming distance:

$$d(x, y) = \#\{i \in [n] : x_i \neq y_i\}$$

- The metric defines a natural group $\mathbb{B}$ consisting of distance-preserving bijections.

Why do Krawtchouk polynomials exist?

- We think of the hypercube $X = \{-1, 1\}^n$ as a metric space with the Hamming distance:

$$d(x, y) = \#\{i \in [n] : x_i \neq y_i\}$$

- The metric defines a natural group $\mathbb{B}$ consisting of distance-preserving bijections.On the hypercube this group is generated by permutations and sign changes.

Why do Krawtchouk polynomials exist?

- We think of the hypercube $X = \{-1, 1\}^n$ as a metric space with the Hamming distance:

$$d(x, y) = \#\{i \in [n] : x_i \neq y_i\}$$

- The metric defines a natural group $\mathbb{B}$ consisting of distance-preserving bijections.On the hypercube this group is generated by permutations and sign changes.
- If $x_0 = (1, \ldots, 1)$ the subgroup $H \subseteq \mathbb{B}$ of elements fixing x_0 is precisely the permutations.

We thus have a pair group, subgroup $(\mathbb{B}, H)$.

This pair has several miraculous properties:

1. The isotypical decomposition of $\mathbb{R}[X]$ as a $\mathbb{B}$-representation is **of multiplicity one**.

$$\mathbb{R}[X] = W_0 \oplus W_1 \oplus \cdots \oplus W_n$$

2. For any $g(t)$ the map $\Gamma_g : \mathbb{R}[X] \to \mathbb{R}[X]$ is a morphism of representations so behaves **like a multiple λ_i of the identity** in each W_i.
3. Each W_i contains a unique copy of the trivial representation, when seeing as an H-representation (this follows from the Frobenius character formula).

This pair has several miraculous properties:

1. The isotypical decomposition of $\mathbb{R}[X]$ as a $\mathbb{B}$-representation is **of multiplicity one**.

$$\mathbb{R}[X] = W_0 \oplus W_1 \oplus \cdots \oplus W_n$$

2. For any $g(t)$ the map $\Gamma_g : \mathbb{R}[X] \to \mathbb{R}[X]$ is a morphism of representations so behaves **like a multiple λ_i of the identity** in each W_i.
3. Each W_i contains a unique copy of the trivial representation, when seeing as an H-representation (this follows from the Frobenius character formula). **The Krawtchouk polynomials are generators of these spaces and can be recovered from them by suitable normalizations**.

Remark.

The last two properties follow from the first.

This pair has several miraculous properties:

1. The isotypical decomposition of $\mathbb{R}[X]$ as a $\mathbb{B}$-representation is **of multiplicity one**.

$$\mathbb{R}[X] = W_0 \oplus W_1 \oplus \cdots \oplus W_n$$

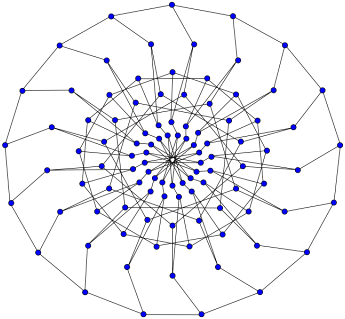
2. For any $g(t)$ the map $\Gamma_g : \mathbb{R}[X] \to \mathbb{R}[X]$ is a morphism of representations so behaves **like a multiple λ_i of the identity** in each W_i.
3. Each W_i contains a unique copy of the trivial representation, when seeing as an H-representation (this follows from the Frobenius character formula). **The Krawtchouk polynomials are generators of these spaces and can be recovered from them by suitable normalizations**.

Remark.

The last two properties follow from the first. A pair $(\mathbb{B}, H)$ with the first property is called a **Gelfand pair**.

Hypercubes in the multiverse

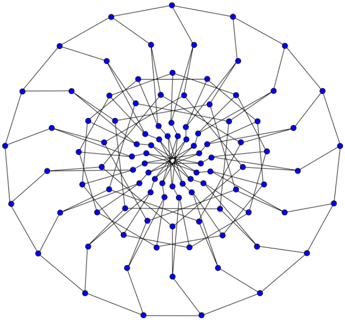
Let X be a finite metric space and let $\mathbb{B}$ be the group of distance-preserving bijections. Fix $x_0 \in X$ and let $H := \mathrm{Stab}(x_0)$.



Definition.

A finite metric space X is **doubly-transitive** *if for any $x_1, x_2, y_1, y_2 \in X$ with $d(x_1, x_2) = d(y_1, y_2)$ there exists an element $g \in \mathbb{B}$ with $y_1 = gx_1$ and $y_2 = gx_2$.*

Hypercubes in the multiverse



There are at least 19 infinite families of doubly transitive graphs, including **hypercubes**, *Cocktail party graphs, Johnson graphs, Grassmann graphs, Paley graphs, etc.*

Hypercubes in the multiverse

For a finite metric space X let $\mathbb{B} := \mathrm{Aut}(X)$ and $n := \mathrm{diam}(\mathrm{X})$.

Theorem. (-)

If X is doubly transitive then the following statements hold:

1. $\mathbb{R}[X]^H = \mathbb{R}[d(x_0, x)] = \mathbb{R}[\ell]/\prod_{j \in \mathrm{range}(d)}(\ell - j)$.
2. $\mathbb{R}[X]$ *decomposes into* $\mathbb{B}$*-irreducibles* W_j *in a multiplicity free manner and every* $\mathbb{B}$*-irreducible contains a unique copy of the* H*-trivial representation.*
3. *There are unique univariate polynomials* $\hat{K}_j(t)$ *such that* $\hat{K}_j(d(x,y))$ *is the Christoffel-Darboux kernel in* W_j.
4. *There is an embedding* $X \subseteq \mathbb{R}^e$ *uniquely specified up to orthogonal transformations. The speed of convergence of the SOS hierarchy on this embedding is bounded by a univariate SDP using the* $\hat{K}_j(t)$.