Kernel sums of squares for optimization and beyond

Francis Bach

INRIA - Ecole Normale Supérieure, Paris, France

Joint work with Ulysse Marteau-Ferey, Alessandro Rudi, Eloise Berthier, Justin Carpentier, Boris Muzellec, Adrien Vacher, François-Xavier Vialard

Rentrée des Masters, FMJH - August 31, 2022

Inference in “high” dimension

- **Convex optimization for machine learning**
  - Linear models
  - dimension $d = 10^{12}$

Inference in “high” dimension

- **Convex optimization for machine learning**
  - Linear models
  - dimension $d = 10^{12}$
- **Optimization for (deep) neural networks**
  - Non-linear models with strong prior structure
  - dimension $d = 10^6$, number of parameters $d = 10^9$

Inference in "high" dimension

- **Convex optimization for machine learning**
  - Linear models
  - dimension $d = 10^{12}$
- **Optimization for (deep) neural networks**
  - Non-linear models with strong prior structure
  - dimension $d = 10^6$, number of parameters $d = 10^9$
- **Curse of dimensionality**
  - Non-linear problems without strong prior structure
  - $d = 10$ or 100
  - Global optimization, optimal transport, optimal control, etc.

Inference in "high" dimension

- **Convex optimization for machine learning**
  - Linear models
  - dimension $d = 10^{12}$
- **Optimization for (deep) neural networks**
  - Non-linear models with strong prior structure
  - dimension $d = 10^6$, number of parameters $d = 10^9$
- **Curse of dimensionality**
  - Non-linear problems without strong prior structure
  - $d = 10$ or 100
  - Global optimization, optimal transport, optimal control, etc.
  - Representation of non-negative functions

Global optimization

- **Zero-th order minimization**

$$\min_{x \in \Omega} f(x)$$

  - $\Omega \subset \mathbb{R}^d$ simple compact subset (e.g., $[-1, 1]^d$)
  - f with some bounded derivatives
  - access to function values

Global optimization

- **Zero-th order minimization**

$$\min_{x \in \Omega} f(x)$$

  - $\Omega \subset \mathbb{R}^d$ simple compact subset (e.g., $[-1, 1]^d$)
  - f with some bounded derivatives
  - access to function values

- **No convexity assumption**

Global optimization

- **Zero-th order minimization**

$$\min_{x \in \Omega} f(x)$$

  - $\Omega \subset \mathbb{R}^d$ simple compact subset (e.g., $[-1,1]^d$)
  - f with some bounded derivatives
  - access to function values

- **No convexity assumption**

- **Many applications**
  - Hyperparameter optimization in machine learning
  - Industry

Optimal algorithms

- **Goal**: Find $\hat{x} \in \Omega$ such that $f(\hat{x}) - \min_{x \in \Omega} f(x) \leqslant \varepsilon$
  - Lowest number of function calls
  - Worst-case guarantees over all functions f in some convex set $\mathcal{F}$

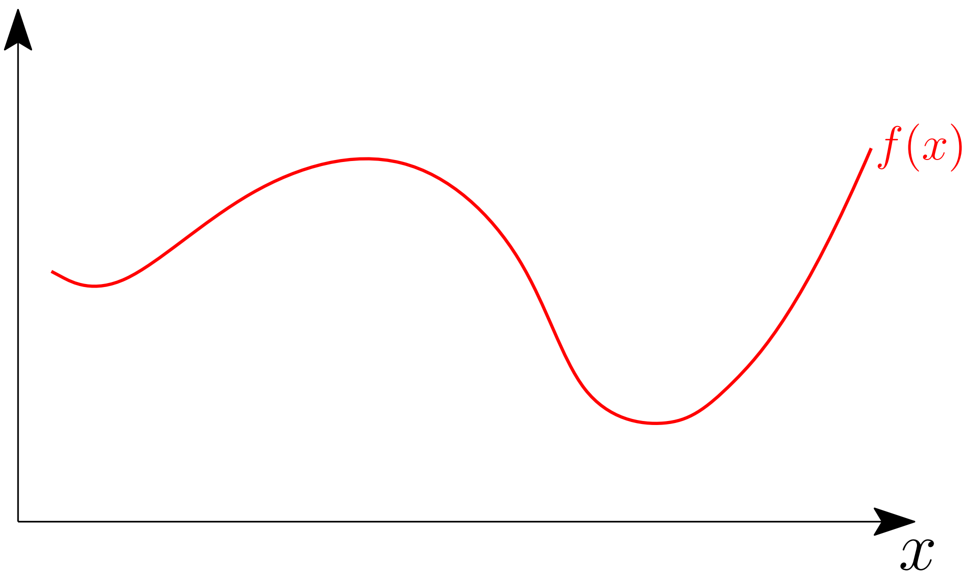
$$\sup_{f \in \mathcal{F}} \Big\{ f(\hat{x}) - \min_{x \in \Omega} f(x) \Big\} \leqslant \varepsilon$$

Optimal algorithms

- **Goal**: Find $\hat{x} \in \Omega$ such that $f(\hat{x}) - \min_{x \in \Omega} f(x) \leqslant \varepsilon$
  - Lowest number of function calls
  - Worst-case guarantees over all functions f in some convex set $\mathcal{F}$

$$\sup_{f \in \mathcal{F}} \Big\{ f(\hat{x}) - \min_{x \in \Omega} f(x) \Big\} \leqslant \varepsilon$$

- **Equivalence to uniform function approximation** (Novak, 2006)
  - Simplest algorithm: approximate f by $\hat{f}$ and minimize $\hat{f}$

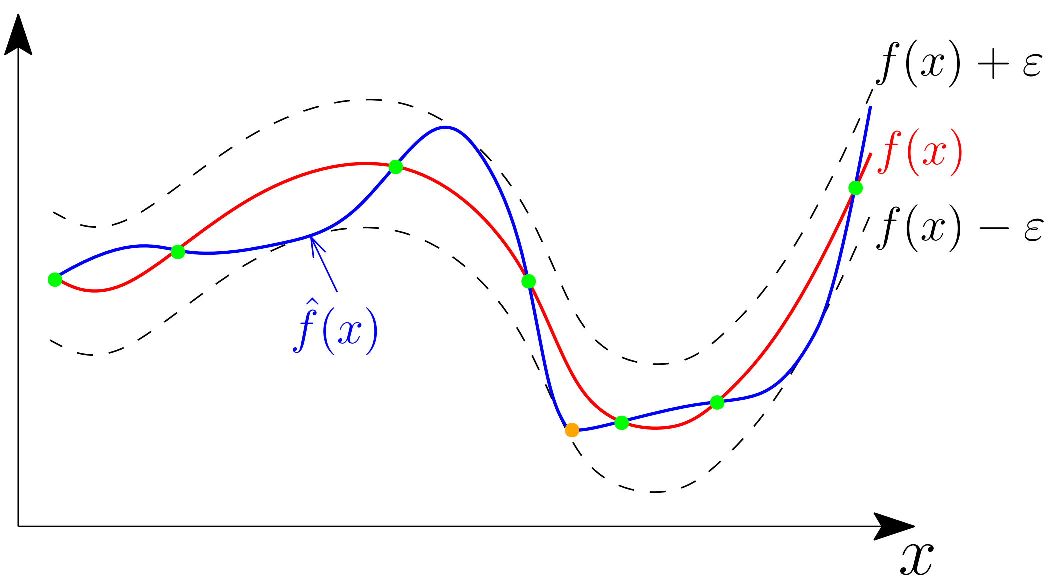


Optimal algorithms

- **Goal**: Find $\hat{x} \in \Omega$ such that $f(\hat{x}) - \min_{x \in \Omega} f(x) \leqslant \varepsilon$
  - Lowest number of function calls
  - Worst-case guarantees over all functions f in some convex set $\mathcal{F}$

$$\sup_{f \in \mathcal{F}} \left\{ f(\hat{x}) - \min_{x \in \Omega} f(x) \right\} \leqslant \varepsilon$$

- **Equivalence to uniform function approximation** (Novak, 2006)
  - Simplest algorithm: approximate f by $\hat{f}$ and minimize $\hat{f}$



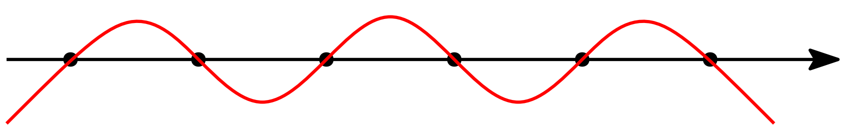
Optimal rates

- **Optimal worst-case performance over** $\mathcal{F}$ (Novak, 2006)
  - $n =$ number of function evaluations
  - $\mathcal{F} =$ Lipschitz-continuous functions: $n \propto \varepsilon^{-d}$

Optimal rates

- **Optimal worst-case performance over $\mathcal{F}$** (Novak, 2006)
  - $n =$ number of function evaluations
  - $\mathcal{F} =$ Lipschitz-continuous functions: $n \propto \varepsilon^{-d}$
  - $\mathcal{F} = m$ bounded derivatives: $n \propto \varepsilon^{-d/m}$

- **Smoothness to circumvent the curse of dimensionality**
  - NB: constants may depend (exponentially) in d



Optimal rates

- **Optimal worst-case performance over $\mathcal{F}$** (Novak, 2006)
  - $n =$ number of function evaluations
  - $\mathcal{F} =$ Lipschitz-continuous functions: $n \propto \varepsilon^{-d}$
  - $\mathcal{F} = m$ bounded derivatives: $n \propto \varepsilon^{-d/m}$
- **Smoothness to circumvent the curse of dimensionality**
  - NB: constants may depend (exponentially) in d
- **Algorithms have exponential running-time complexity**
  - Approximate *then* optimize

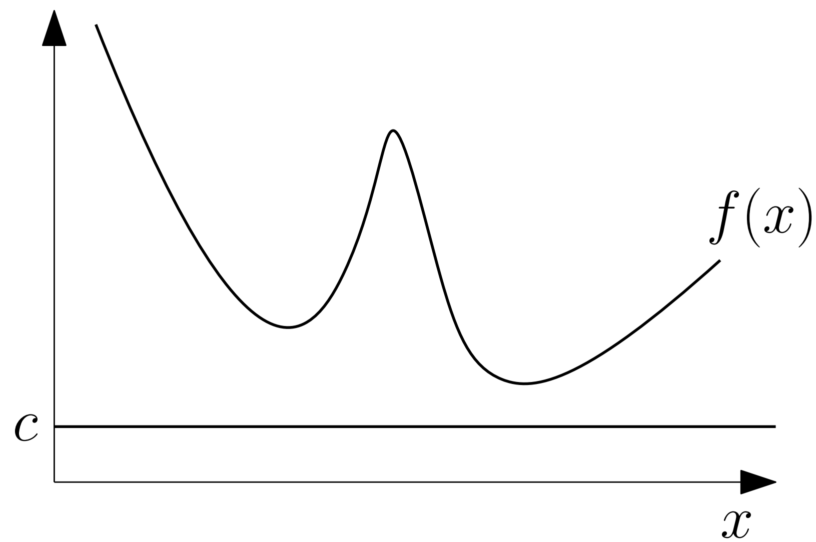
Optimal rates

- **Optimal worst-case performance over $\mathcal{F}$** (Novak, 2006)
  - $n =$ number of function evaluations
  - $\mathcal{F} =$ Lipschitz-continuous functions: $n \propto \varepsilon^{-d}$
  - $\mathcal{F} = m$ bounded derivatives: $n \propto \varepsilon^{-d/m}$
- **Smoothness to circumvent the curse of dimensionality**
  - NB: constants may depend (exponentially) in d
- **Algorithms have exponential running-time complexity**
  - Approximate *then* optimize
- **Algorithms with polynomial-time complexity in n?**
  - Approximate *and* optimize

Reformulations

- **Equivalent convex problem**

$$\min_{x \in \Omega} f(x) = \sup_{c \in \mathbb{R}} c \quad \text{such that} \quad \forall x \in \Omega,\ f(x) - c \geqslant 0$$

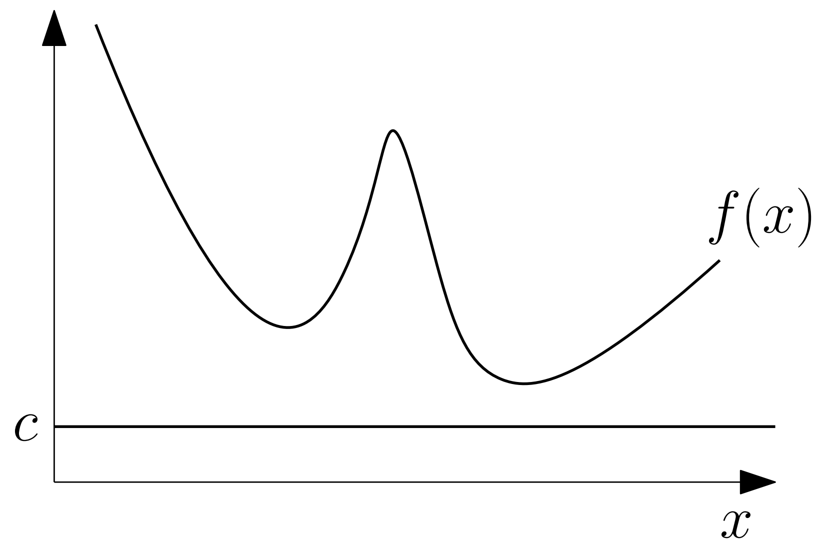


  - All optimization problems are convex!

Reformulations

- **Equivalent convex problem**

$$\min_{x\in\Omega} f(x) = \sup_{c\in\mathbb{R}} c \quad \text{such that} \quad \forall x \in \Omega,\ f(x) - c \geqslant 0$$



  – All optimization problems are convex!

- **Need to represent non-negative functions** (such as $f(x) - c$)

Representing non-negative functions

- **Assumption**: $g(x)$ can be represented as $g(x) = \langle \phi(x), G\phi(x) \rangle$
  - with G symmetric operator
  - Assume constant function can be represented as $1 = \langle u, \phi(x) \rangle$
  - Example: set of polynomials of degree $2r$
    with $\phi(x)$ composed of monomials of degree r, of dimension $\binom{d+r}{r}$

Representing non-negative functions

- **Assumption**: $g(x)$ can be represented as $g(x) = \langle \phi(x), G\phi(x) \rangle$
  - with G symmetric operator
  - Assume constant function can be represented as $1 = \langle u, \phi(x) \rangle$
  - Example: set of polynomials of degree $2r$
    with $\phi(x)$ composed of monomials of degree r, of dimension $\binom{d+r}{r}$

- **Positivity through "sums-of-squares"**
  - If $G \succcurlyeq 0$, then $\forall x \in \Omega,\ g(x) = \langle \phi(x), G\phi(x) \rangle \geqslant 0$
  - Then, $g(x) = \sum_{i \in I} \lambda_i \langle \phi(x), (h_i \otimes h_i)\phi(x) \rangle = \sum_{i \in I} \lambda_i \langle \phi(x), h_i \rangle^2$

Representing non-negative functions

- **Assumption**: $g(x)$ can be represented as $g(x) = \langle \phi(x), G\phi(x) \rangle$
  - with G symmetric operator
  - Assume constant function can be represented as $1 = \langle u, \phi(x) \rangle$
  - Example: set of polynomials of degree $2r$
    with $\phi(x)$ composed of monomials of degree r, of dimension $\binom{d+r}{r}$

- **Positivity through "sums-of-squares"**
  - If $G \succcurlyeq 0$, then $\forall x \in \Omega,\ g(x) = \langle \phi(x), G\phi(x) \rangle \geqslant 0$
  - Then, $g(x) = \sum_{i \in I} \lambda_i \langle \phi(x), (h_i \otimes h_i)\phi(x) \rangle = \sum_{i \in I} \lambda_i \langle \phi(x), h_i \rangle^2$

- **Are all non-negative functions sums-of-squares?**
  - Polynomials: no if $d > 1$ (see, e.g., Rudin, 2000)

Global optimization with sums of square polynomials

- **Replace** $f(x) - c \geqslant 0$ by $f(x) = c + \langle \phi(x), A\phi(x) \rangle$ with $A \succcurlyeq 0$
  - represented as $F = c \cdot u \otimes u + A$

Global optimization with sums of square polynomials

- **Replace** $f(x) - c \geqslant 0$ by $f(x) = c + \langle \phi(x), A\phi(x) \rangle$ with $A \succcurlyeq 0$
  - represented as $F = c \cdot u \otimes u + A$

- **Sum-of-squares optimization** (Lasserre, 2001; Parrilo, 2003)

$$\sup_{c \in \mathbb{R},\ A \succcurlyeq 0} c \quad \text{such that} \quad \forall x \in \mathbb{R}^d,\ f(x) = c + \langle \phi(x), A\phi(x) \rangle$$

  - Equivalent to original problem if $f(x) - f_*$ is a sum-of-squares

Global optimization with sums of square polynomials

- **Replace** $f(x) - c \geqslant 0$ by $f(x) = c + \langle \phi(x), A\phi(x) \rangle$ with $A \succcurlyeq 0$
  - represented as $F = c \cdot u \otimes u + A$

- **Sum-of-squares optimization** (Lasserre, 2001; Parrilo, 2003)

$$\sup_{c \in \mathbb{R},\ A \succcurlyeq 0} c \quad \text{such that} \quad \forall x \in \mathbb{R}^d,\ f(x) = c + \langle \phi(x), A\phi(x) \rangle$$

  - Equivalent to original problem if $f(x) - f_*$ is a sum-of-squares
  - If not, and if localization set $\Omega = \{x, \|x\|^2 \leqslant R^2\}$ is known,

$$\forall x \in \Omega,\ f(x) > 0 \quad \Rightarrow \quad \forall x \in \mathbb{R}^d,\ f(x) = q(x) + (R^2 - \|x\|^2)p(x)$$

    with p and q sums-of-squares polynomials (of unknown degree)
  - Needs "hierarchies"

Representing more general functions with RKHSs

- **Reproducing Kernel Hilbert Space (RKHS)** :
  - Hilbert space of functions $g \in \mathcal{H}, \ g : \mathbb{R}^d \to \mathbb{R}$
  - Representation as linear form : $g(x) = \langle g, \phi(x) \rangle$
  - Kernel : $k(x, x') = \langle \phi(x), \phi(x') \rangle$ (computable)

Representing more general functions with RKHSs

- **Reproducing Kernel Hilbert Space (RKHS)** :
  - Hilbert space of functions $g \in \mathcal{H}$, $g : \mathbb{R}^d \to \mathbb{R}$
  - Representation as linear form : $g(x) = \langle g, \phi(x) \rangle$
  - Kernel : $k(x, x') = \langle \phi(x), \phi(x') \rangle$ (computable)

- **Example : Sobolev spaces** (Berlinet and Thomas-Agnan, 2011)
  - Sobolev spaces $H^s(\Omega)$ with $\Omega \subset \mathbb{R}^d$, $s > d/2$

$$\langle f, g \rangle = \sum_{|\alpha| \leq s} \int_\Omega \partial^\alpha f(x) \cdot \partial^\alpha g(x) dx$$

  - Example $s = d/2 + 1/2$: $k(x, y) = \exp(-\|x - y\|)$

Representing more general functions with RKHSs

- **Reproducing Kernel Hilbert Space (RKHS)** :
  - Hilbert space of functions $g \in \mathcal{H}, \; g : \mathbb{R}^d \to \mathbb{R}$
  - Representation as linear form : $g(x) = \langle g, \phi(x) \rangle$
  - Kernel : $k(x, x') = \langle \phi(x), \phi(x') \rangle$ (computable)

- **Everything can be expressed using only the kernel function k**
  - Useful when dealing with function evaluations
  - Representer theorem (Kimeldorf and Wahba, 1971): Minimizing $L(g(x_1), \dots, g(x_n)) + \frac{\lambda}{2}\|g\|^2$ can be done by restricting to

$$g(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i)$$

  - Then $g(x_j) = \sum_{i=1}^{n} \alpha_i k(x_j, x_i)$ and $\|g\|^2 = \sum_{i,j=1}^{n} \alpha_i \alpha_j k(x_i, x_j)$

Going infinite-dimensional (Rudi, Marteau-Ferey, and Bach, 2020)

$$\sup_{c \in \mathbb{R},\ A \succcurlyeq 0} \quad c \quad \text{such that} \quad \forall x \in \Omega,\ f(x) = c + \langle \phi(x), A\phi(x) \rangle$$

- $\phi(x) \in \mathcal{H}$ Hilbert space so that $\langle w, \phi(x) \rangle$ spans a Sobolev space
  - $s > d/2$ squared-integrable derivative
  - Reproducing kernel Hilbert space (RKHS)
  - $k(x, y) = \langle \phi(x), \phi(y) \rangle = \exp(-\|x - y\|)$ for $s = d/2 + 1/2$.
  - See, e.g., Berlinet and Thomas-Agnan (2011)

Going infinite-dimensional (Rudi, Marteau-Ferey, and Bach, 2020)

$$\sup_{c \in \mathbb{R},\ A \succcurlyeq 0} \ c \quad \text{such that} \quad \forall x \in \Omega,\ f(x) = c + \langle \phi(x), A\phi(x) \rangle$$

- $\phi(x) \in \mathcal{H}$ Hilbert space so that $\langle w, \phi(x) \rangle$ spans a Sobolev space
  - $s > d/2$ squared-integrable derivative
  - Reproducing kernel Hilbert space (RKHS)
  - $k(x, y) = \langle \phi(x), \phi(y) \rangle = \exp(-\|x - y\|)$ for $s = d/2 + 1/2$.
  - See, e.g., Berlinet and Thomas-Agnan (2011)

- **Theorem**: $\exists A_* \succcurlyeq 0$ such that $\forall x \in \Omega,\ f(x) = f_* + \langle \phi(x), A_*\phi(x) \rangle$
  - If f has isolated strict-second order minima in $\mathring{\Omega}$, and f is $(s + 3)$-times differentiable

Going infinite-dimensional (Rudi, Marteau-Ferey, and Bach, 2020)

$$\sup_{c \in \mathbb{R},\ A \succcurlyeq 0} \ c \quad \text{such that} \quad \forall x \in \Omega,\ f(x) = c + \langle \phi(x), A\phi(x) \rangle$$

- $\phi(x) \in \mathcal{H}$ Hilbert space so that $\langle w, \phi(x) \rangle$ spans a Sobolev space
  - $s > d/2$ squared-integrable derivative
  - Reproducing kernel Hilbert space (RKHS)
  - $k(x, y) = \langle \phi(x), \phi(y) \rangle = \exp(-\|x - y\|)$ for $s = d/2 + 1/2$.
  - See, e.g., Berlinet and Thomas-Agnan (2011)

- **Theorem**: $\exists A_* \succcurlyeq 0$ such that $\forall x \in \Omega,\ f(x) = f_* + \langle \phi(x), A_*\phi(x) \rangle$
  - If f has isolated strict-second order minima in $\mathring{\Omega}$, and f is $(s+3)$-times differentiable

$\Rightarrow$ Equivalent to original problem, but infinite-dimensional

Controlled approximation through sampling

- **Subsample n points $x_1, \dots, x_n \in \Omega$ and solve**

$$\sup_{c \in \mathbb{R},\ A \succcurlyeq 0} c - \lambda \operatorname{tr}(A) \ \text{ such that } \ \forall i \in \{1, \dots, n\},\ f(x_i) = c + \langle \phi(x_i), A\phi(x_i) \rangle$$

Controlled approximation through sampling

- **Subsample n points $x_1, \dots, x_n \in \Omega$ and solve**

$$\sup_{c \in \mathbb{R},\ A \succcurlyeq 0} c - \lambda \operatorname{tr}(A) \text{ such that } \forall i \in \{1, \dots, n\},\ f(x_i) = c + \langle \phi(x_i), A\phi(x_i) \rangle$$

- **Approximation guarantees** (Rudi, Marteau-Ferey, and Bach, 2020)
  - With random samples, $n \approx \varepsilon^{-d/(m-d/2-3)}$ (up to logarithmic terms)
  - To be compared to optimal rate $n \approx \varepsilon^{-d/(m-d/2)}$
  - Constraint $m \geqslant \frac{d}{2} + 3$ can be lifted

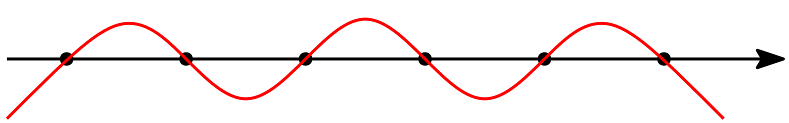
Controlled approximation through sampling

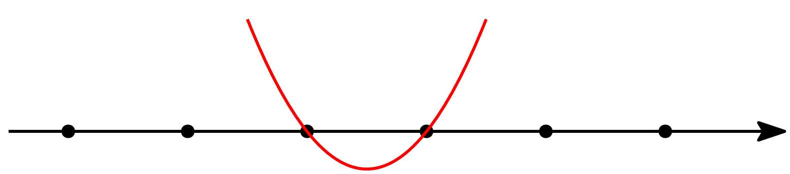
- **Subsample n points $x_1, \dots, x_n \in \Omega$ and solve**

$$\sup_{c \in \mathbb{R},\ A \succcurlyeq 0} \ c - \lambda \operatorname{tr}(A) \ \text{ such that } \ \forall i \in \{1, \dots, n\},\ f(x_i) = c + \langle \phi(x_i), A\phi(x_i) \rangle$$

- **Approximation guarantees** (Rudi, Marteau-Ferey, and Bach, 2020)
  - With random samples, $n \approx \varepsilon^{-d/(m-d/2-3)}$ (up to logarithmic terms)
  - To be compared to optimal rate $n \approx \varepsilon^{-d/(m-d/2)}$
  - Constraint $m \geqslant \frac{d}{2} + 3$ can be lifted

- **Subsampling inequalities as $f(x_i) \geqslant c$ directly?**
  - cannot improve on $n \approx \varepsilon^{-d}$





Controlled approximation through sampling

- **Subsample n points $x_1, \dots, x_n \in \Omega$ and solve**

$$\sup_{c \in \mathbb{R},\ A \succcurlyeq 0} c - \lambda \operatorname{tr}(A) \text{ such that } \forall i \in \{1, \dots, n\},\ f(x_i) = c + \langle \phi(x_i), A\phi(x_i) \rangle$$

- **Finite-dimensional algorithm through representer theorem**
  - Marteau-Ferey, Bach, and Rudi (2020)
  - Restrict optimization to $A = \sum_{i,j=1}^{n} C_{ij} \phi(x_i) \otimes \phi(x_j)$ with $C \succcurlyeq 0$

Controlled approximation through sampling

- **Subsample n points $x_1, \dots, x_n \in \Omega$ and solve**

$$\sup_{c \in \mathbb{R},\ A \succcurlyeq 0} c - \lambda \operatorname{tr}(A) \ \text{ such that } \ \forall i \in \{1, \dots, n\},\ f(x_i) = c + \langle \phi(x_i), A\phi(x_i) \rangle$$

- **Finite-dimensional algorithm through representer theorem**
  - Marteau-Ferey, Bach, and Rudi (2020)
  - Restrict optimization to $A = \sum_{i,j=1}^{n} C_{ij} \phi(x_i) \otimes \phi(x_j)$ with $C \succcurlyeq 0$

- **Semi-definite programming problem**
  - Complexity $O(n^{3.5} \log \frac{1}{\varepsilon})$ by interior point method
  - More efficient Newton algorithm in $O(n^3)$

Final algorithm

- **Input:** $f : \mathbb{R}^d \to \mathbb{R}$, $\Omega \subset \mathbb{R}^d, n \geqslant 0, \lambda > 0, s > d/2$

Final algorithm

- **Input:** $f : \mathbb{R}^d \to \mathbb{R}$, $\Omega \subset \mathbb{R}^d, n \geqslant 0, \lambda > 0, s > d/2$

1. **Sampling:** $\{x_1, \ldots, x_n\}$ sampled i.i.d. uniformly on Ω

Final algorithm

- **Input:** $f : \mathbb{R}^d \to \mathbb{R}$, $\Omega \subset \mathbb{R}^d, n \geqslant 0, \lambda > 0, s > d/2$

1. **Sampling:** $\{x_1, \dots, x_n\}$ sampled i.i.d. uniformly on Ω

2. **Feature computation**

   – Compute $K_{ij} = k(x_i, x_j)$ for k Sobolev kernel of smoothness s
   – Compute square root of $K = R^\top R \in \mathbb{R}^{n \times n}$
   – Set $\Phi_j = j$-th column of R, $\forall j \in \{1, \dots, n\}$
   – Set $f_j = f(x_j)$, $\forall j \in \{1, \dots, n\}$

Final algorithm

- **Input:** $f : \mathbb{R}^d \to \mathbb{R}$, $\Omega \subset \mathbb{R}^d, n \geqslant 0, \lambda > 0, s > d/2$

1. **Sampling:** $\{x_1, \dots, x_n\}$ sampled i.i.d. uniformly on Ω

2. **Feature computation**
   - Compute $K_{ij} = k(x_i, x_j)$ for k Sobolev kernel of smoothness s
   - Compute square root of $K = R^\top R \in \mathbb{R}^{n \times n}$
   - Set $\Phi_j = j$-th column of R, $\forall j \in \{1, \dots, n\}$
   - Set $f_j = f(x_j)$, $\forall j \in \{1, \dots, n\}$

3. **Solve** $\max_{c \in \mathbb{R}, B \succcurlyeq 0} c - \lambda \operatorname{tr}(B)$ s. t. $\forall j \in \{1, \dots, n\}, f_j - c = \Phi_j^\top B \Phi_j$
   - With Lagrange multipliers $\alpha \in \mathbb{R}^n$

Final algorithm

- **Input:** $f : \mathbb{R}^d \to \mathbb{R}$, $\Omega \subset \mathbb{R}^d, n \geqslant 0, \lambda > 0, s > d/2$

1. **Sampling:** $\{x_1, \dots, x_n\}$ sampled i.i.d. uniformly on Ω

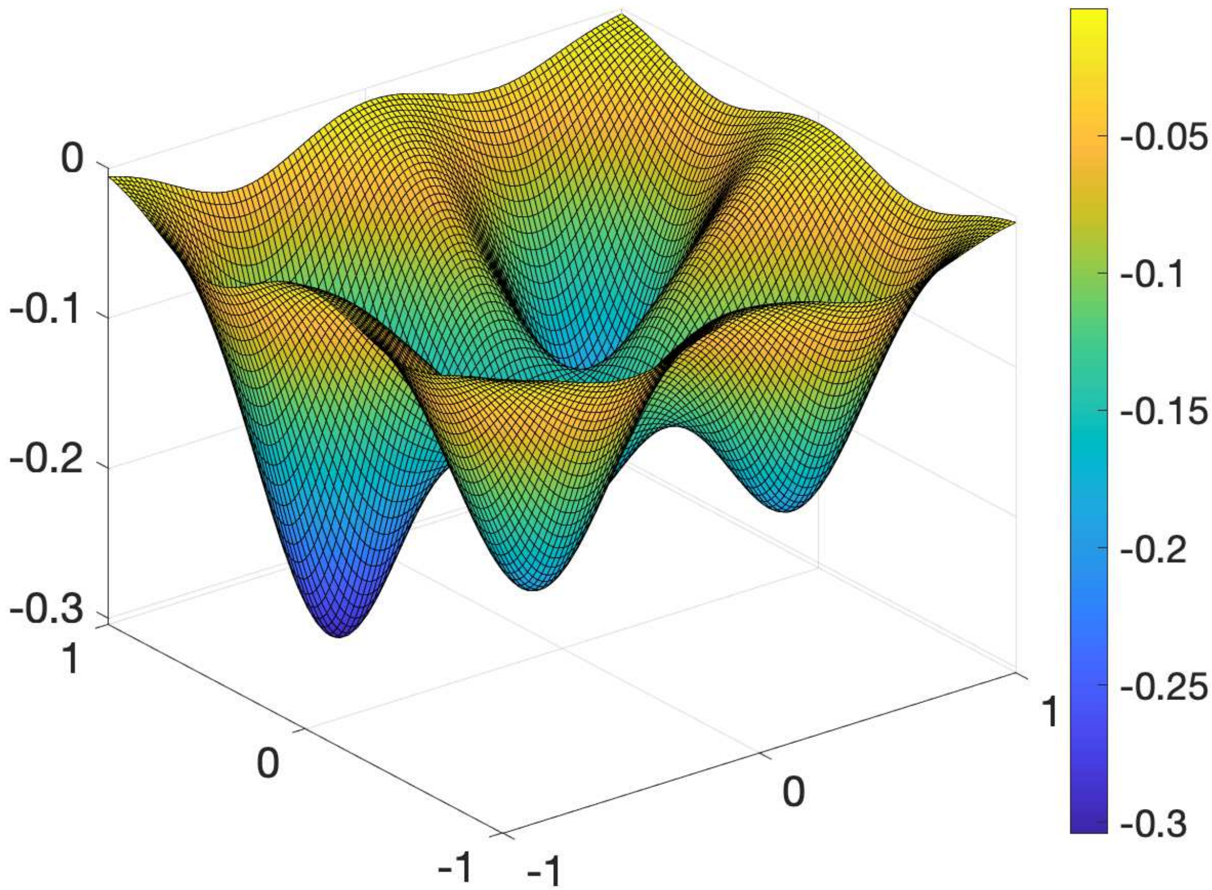
2. **Feature computation**
   - Compute $K_{ij} = k(x_i, x_j)$ for k Sobolev kernel of smoothness s
   - Compute square root of $K = R^\top R \in \mathbb{R}^{n \times n}$
   - Set $\Phi_j = j$-th column of R, $\forall j \in \{1, \dots, n\}$
   - Set $f_j = f(x_j)$, $\forall j \in \{1, \dots, n\}$

3. **Solve** $\max_{c \in \mathbb{R}, B \succcurlyeq 0} c - \lambda \operatorname{tr}(B)$ s. t. $\forall j \in \{1, \dots, n\}, f_j - c = \Phi_j^\top B \Phi_j$
   - With Lagrange multipliers $\alpha \in \mathbb{R}^n$

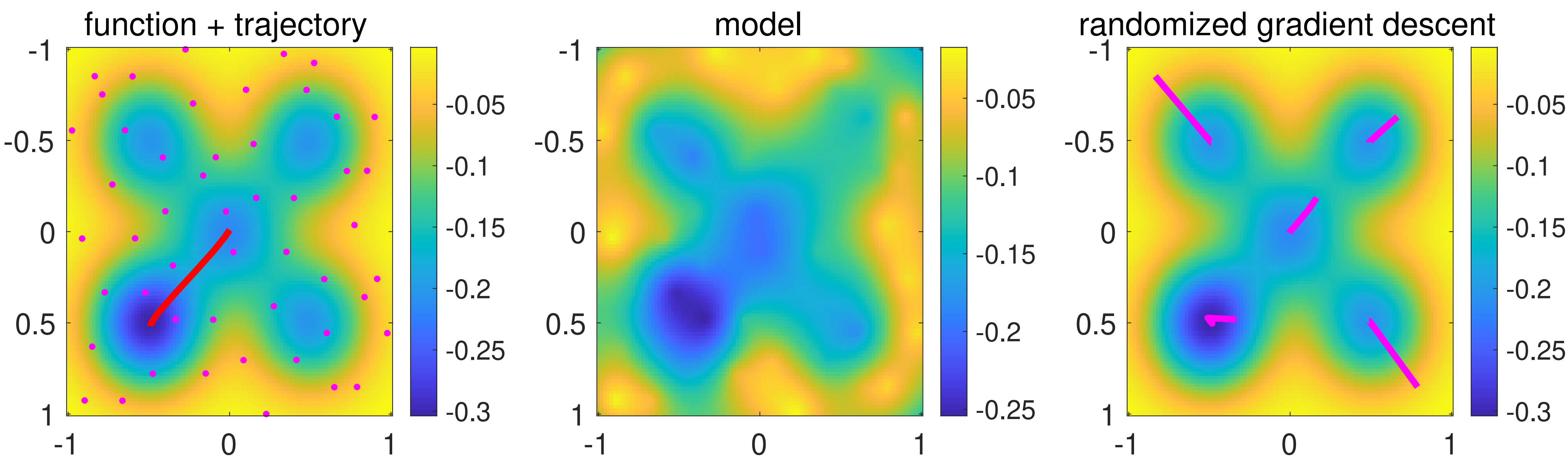
- **Output:** c and $\hat{x} = \sum_{j=1}^n \alpha_j x_j$

Illustration

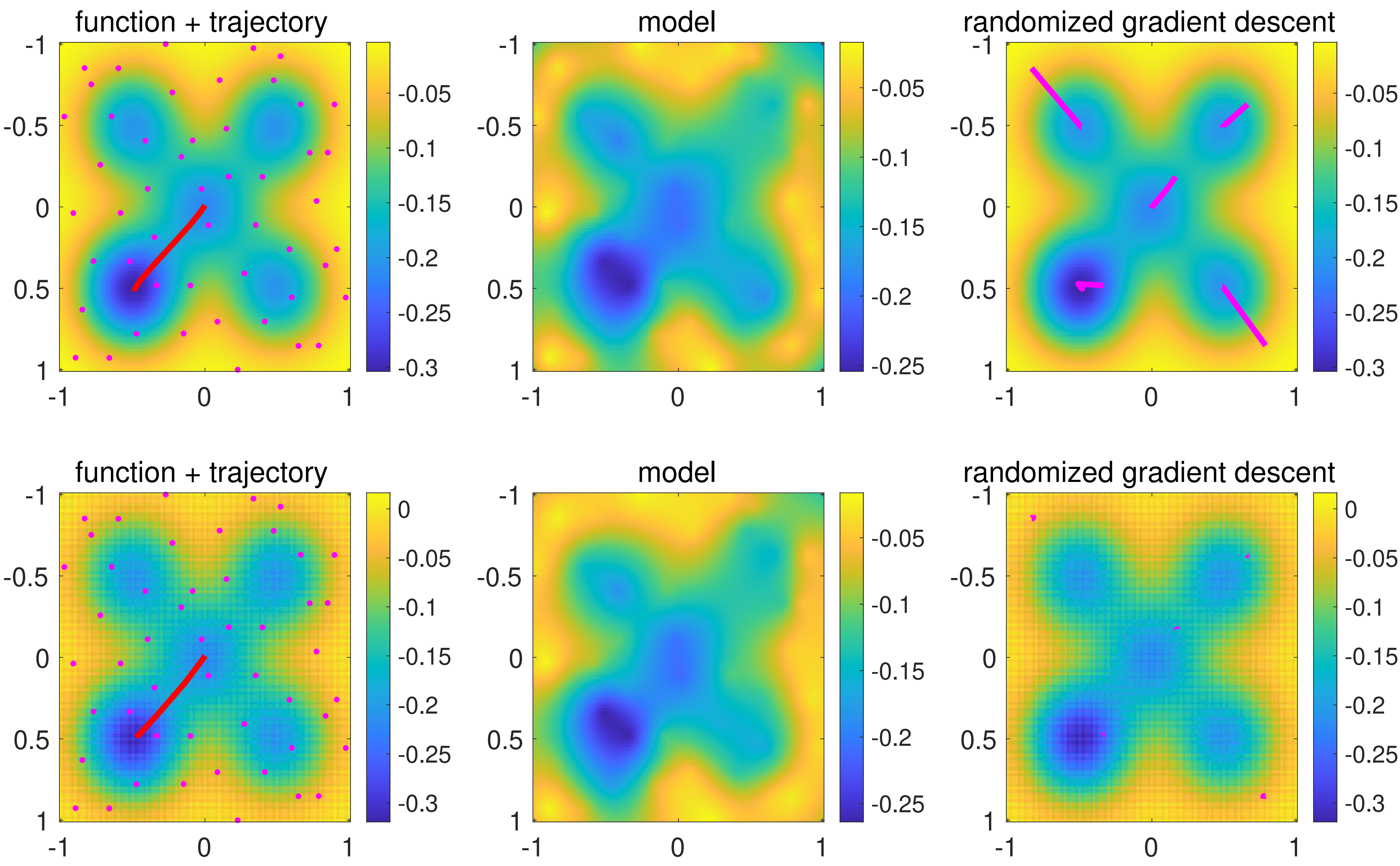
- **Minimization of two-dimensional function**



Illustration

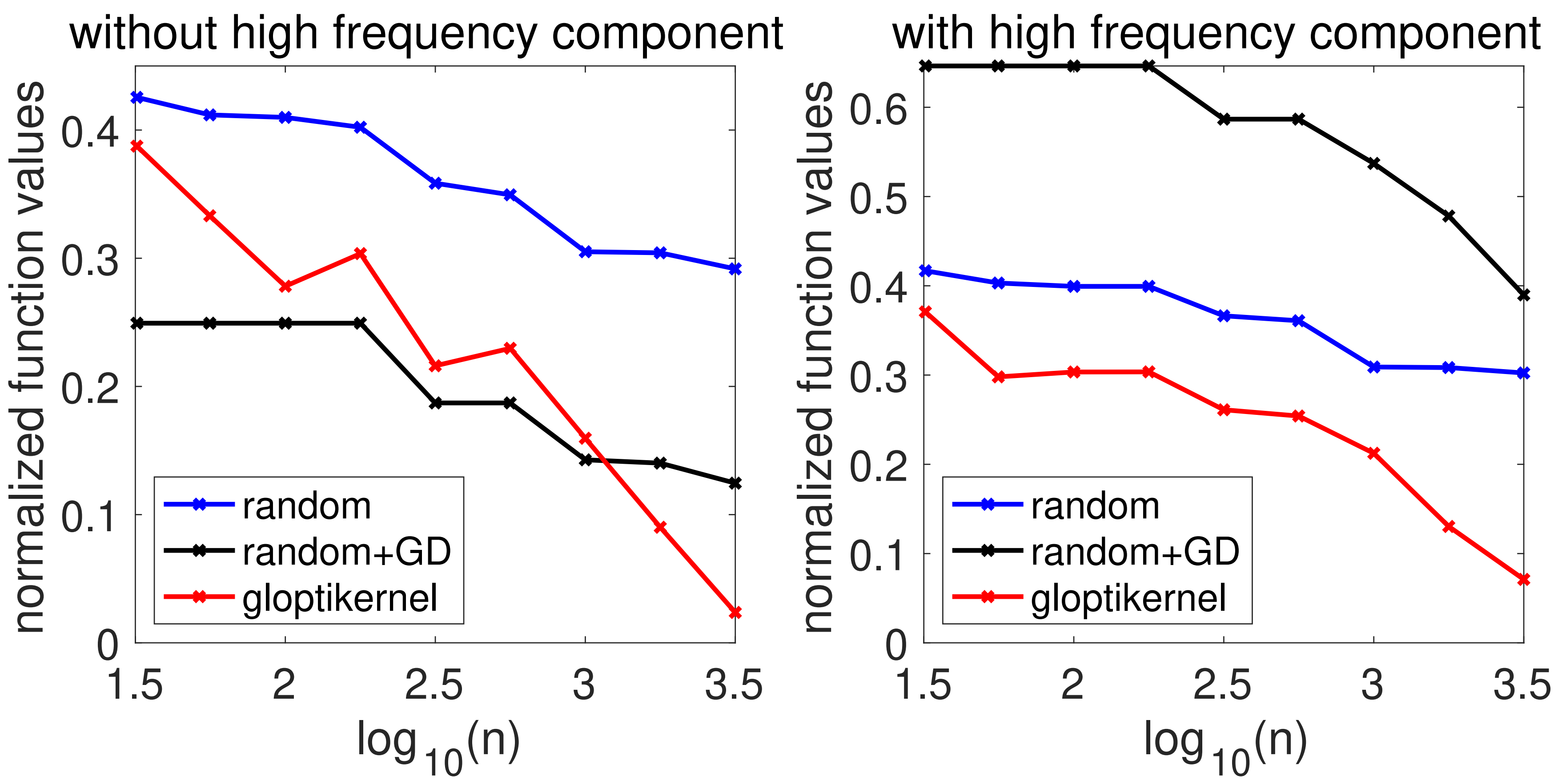


Illustration



Illustration

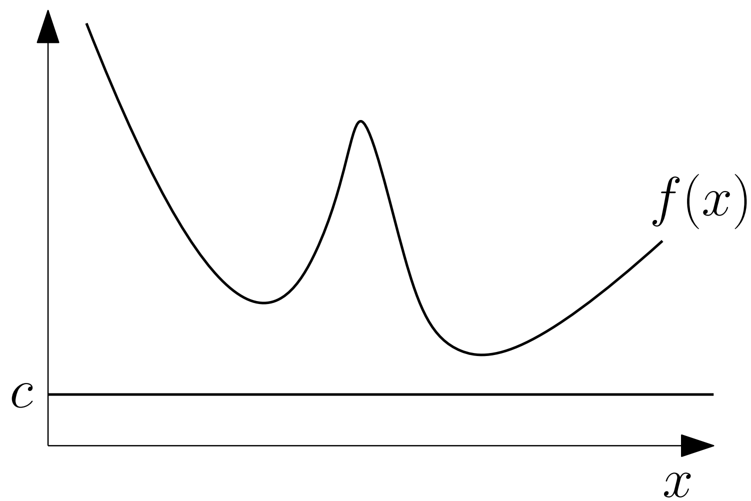
- **Minimization of eight-dimensional function**



Duality

- **Primal problem**

$$\min_{x\in\Omega} f(x) = \sup_{c\in\mathbb{R}} c \quad \text{such that} \quad \forall x \in \Omega,\ f(x) - c \geqslant 0$$



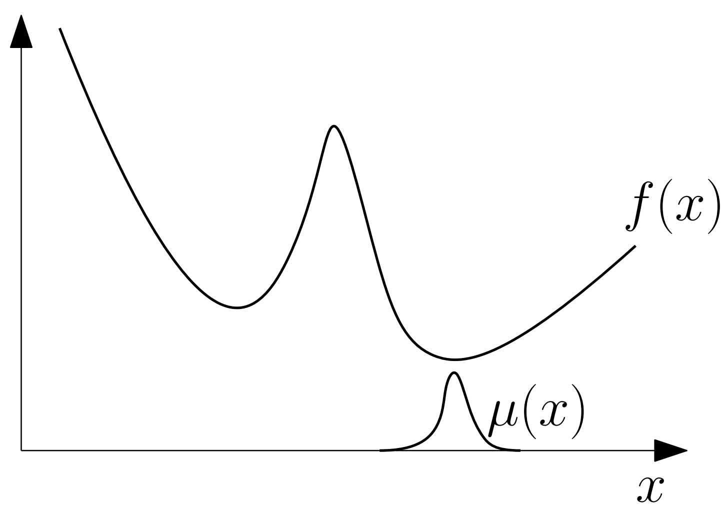
Duality

- **Primal problem**

$$\min_{x\in\Omega} f(x) = \sup_{c\in\mathbb{R}} c \quad \text{such that} \quad \forall x \in \Omega,\ f(x) - c \geqslant 0$$

- **Dual problem on probability measures**

$$\inf_{\mu\in\mathbb{R}^\Omega} \int_\Omega \mu(x)f(x)dx \quad \text{such that} \quad \int_\Omega \mu(x)dx = 1,\ \forall x \in \Omega,\ \mu(x) \geqslant 0$$



Duality with sums-of-squares

- **Primal problem**

$$\min_{x\in\Omega} f(x) = \sup_{c\in\mathbb{R},\ A\succcurlyeq 0} c \quad \text{such that} \quad \forall x \in \Omega,\ f(x) - c = \langle \phi(x), A\phi(x) \rangle$$

- **Dual problem on signed measures**

$$\inf_{\mu\in\mathbb{R}^\Omega} \int_\Omega \mu(x) f(x) dx \quad \text{s. t.} \quad \int_\Omega \mu(x) dx = 1,\ \int_\Omega \mu(x) \phi(x) \otimes \phi(x) \succcurlyeq 0$$

  – Extension of results on polynomials (Lasserre, 2020)

Extension

- **Generic constrained optimization problem**

$$\inf_{\theta \in \Theta} F(\theta) \quad \text{such that} \quad \forall x \in \Omega,\ g(\theta, x) \geqslant 0$$

Extension

- **Generic constrained optimization problem**

$$\inf_{\theta \in \Theta} F(\theta) \quad \text{such that} \quad \forall x \in \Omega,\ g(\theta, x) \geqslant 0$$

- **Sums-of-squares reformulation**

$$\inf_{\theta \in \Theta,\ A \succcurlyeq 0} F(\theta) \quad \text{such that} \quad \forall x \in \Omega,\ g(\theta, x) = \langle \phi(x), A\phi(x) \rangle$$

  – Requires penalization by $\operatorname{tr}(A)$ and subsampling
  – Need representation as sums-of-squares to benefit from smoothness
  – Can be done in the primal or the dual

Extension

- **Generic constrained optimization problem**

$$\inf_{\theta \in \Theta} F(\theta) \quad \text{such that} \quad \forall x \in \Omega, \ g(\theta, x) \geqslant 0$$

- **Sums-of-squares reformulation**

$$\inf_{\theta \in \Theta, \ A \succcurlyeq 0} F(\theta) \quad \text{such that} \quad \forall x \in \Omega, \ g(\theta, x) = \langle \phi(x), A\phi(x) \rangle$$

  – Requires penalization by $\mathrm{tr}(A)$ and subsampling
  – Need representation as sums-of-squares to benefit from smoothness
  – Can be done in the primal or the dual

- **Application to optimal transport and optimal control**

Smooth optimal transport

- **Primal formulation**: $\displaystyle \inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y) d\gamma(x,y)$
  - $\Gamma(\mu,\nu)$ set of probability distributions with marginals μ and ν
- **Dual formulation**: $\displaystyle \sup_{u,v \in C(\mathbb{R}^d)} \int_{\mathcal{X}} u(x) d\mu(x) + \int_{\mathcal{Y}} v(y) d\mu(y)$

  such that $\quad \forall (x,y) \in \mathcal{X} \times \mathcal{Y},\ c(x,y) - u(x) + v(y) \geqslant 0$

Smooth optimal transport

- **Primal formulation**: $\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y) d\gamma(x,y)$
  - $\Gamma(\mu,\nu)$ set of probability distributions with marginals μ and ν

- **Dual formulation**: $\sup_{u,v \in C(\mathbb{R}^d)} \int_{\mathcal{X}} u(x) d\mu(x) + \int_{\mathcal{Y}} v(y) d\mu(y)$

  such that $\forall (x,y) \in \mathcal{X}\times\mathcal{Y},\ c(x,y) - u(x) + v(y) \geqslant 0$

- **Estimation from i.i.d. samples from smooth densities for μ and ν**
  - Rate: from $O(n^{-1/d})$ to $O(n^{-m/d})$ (Weed and Berthet, 2019)
  - No polynomial-time algorithm

Smooth optimal transport

- **Primal formulation**: $\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{Y}} c(x,y) d\gamma(x,y)$
  - $\Gamma(\mu,\nu)$ set of probability distributions with marginals μ and ν

- **Dual formulation**: $\sup_{u,v \in C(\mathbb{R}^d)} \int_{\mathcal{X}} u(x) d\mu(x) + \int_{\mathcal{Y}} v(y) d\mu(y)$

  such that $\forall (x,y) \in \mathcal{X} \times \mathcal{Y},\ c(x,y) - u(x) + v(y) \geqslant 0$

- **Estimation from i.i.d. samples from smooth densities for μ and ν**
  - Rate: from $O(n^{-1/d})$ to $O(n^{-m/d})$ (Weed and Berthet, 2019)
  - No polynomial-time algorithm

- **Kernel sums of squares**: "polynomial"-time algorithm
  - Vacher, Muzellec, Rudi, Bach, and Vialard (2021)

Optimal control / reinforcement learning

- **Optimal control** (Liberzon, 2011)

$$V^*(t_0, x_0) = \inf_{u:[t_0,T]\to\mathcal{U}} \int_{t_0}^{T} L(t, x(t), u(t))\mathsf{d}t + M(x(T))$$

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

Optimal control / reinforcement learning

- **Optimal control** (Liberzon, 2011)

$$V^*(t_0, x_0) = \inf_{u:[t_0,T]\to\mathcal{U}} \int_{t_0}^{T} L(t, x(t), u(t))\mathrm{d}t + M(x(T))$$
$$\forall t \in [t_0, T],\ \dot{x}(t) = f(t, x(t), u(t)), \quad x(t_0) = x_0.$$

- Subsolution of **Hamilton-Jacobi-Bellman** equation (Vinter, 1993)

$$\sup_{V:[0,T]\times\mathcal{X}\to\mathbb{R}} \int V(0, x_0)\mathrm{d}\mu_0(x_0)$$
$$\forall (t,x,u), \quad \frac{\partial V}{\partial t}(t,x) + L(t,x,u) + \nabla V(t,x)^\top f(t,x,u) \geqslant 0$$
$$\forall x, \quad V(T,x) = M(x).$$

Optimal control / reinforcement learning

- Subsolution of **Hamilton-Jacobi-Bellman** equation (Vinter, 1993)

$$\sup_{V:[0,T]\times\mathcal{X}\to\mathbb{R}} \int V(0,x_0)\mathsf{d}\mu_0(x_0)$$

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

$$\forall x, \quad V(T,x) = M(x).$$

Optimal control / reinforcement learning

- Subsolution of **Hamilton-Jacobi-Bellman** equation (Vinter, 1993)

$$\sup_{V:[0,T]\times\mathcal{X}\to\mathbb{R}} \int V(0,x_0)\mathsf{d}\mu_0(x_0)$$

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

$$\forall x, \quad V(T,x) = M(x).$$

- **Polynomial sums-of-squares**
  - Lasserre, Henrion, Prieur, and Trélat (2008)

Optimal control / reinforcement learning

- Subsolution of **Hamilton-Jacobi-Bellman** equation (Vinter, 1993)

$$\sup_{V:[0,T]\times\mathcal{X}\to\mathbb{R}} \int V(0,x_0)\mathsf{d}\mu_0(x_0)$$

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

$$\forall x, \quad V(T,x) = M(x).$$

- **Polynomial sums-of-squares**
  - Lasserre, Henrion, Prieur, and Trélat (2008)
- **Extension to kernel sums-of-squares**
  - Berthier, Carpentier, Rudi, and Bach (2021)
  - Allows some form of modelling

Conclusion

- **Global optimization through kernel approximations**
  - Joint optimization and approximation
  - infinite-dimensional sums-of-squares representation
  - Controlled subsampling with guarantees

Conclusion

- **Global optimization through kernel approximations**
  - Joint optimization and approximation
  - infinite-dimensional sums-of-squares representation
  - Controlled subsampling with guarantees

- **Further improvements**
  - Efficient algorithms below $O(n^3)$ complexity
  - Adaptive choice of sampling points
  - Certificates of optimality

Conclusion

- **Global optimization through kernel approximations**
  - Joint optimization and approximation
  - infinite-dimensional sums-of-squares representation
  - Controlled subsampling with guarantees

- **Further improvements**
  - Efficient algorithms below $O(n^3)$ complexity
  - Adaptive choice of sampling points
  - Certificates of optimality

- **Further extensions**
  - Other infinite-dimensional convex optimization problems?
  - Bayesian inference (Rudi and Ciliberto, 2021; Marteau-Ferey, Rudi, and Bach, 2021)

References

Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer Science & Business Media, 2011.

Eloïse Berthier, Justin Carpentier, Alessandro Rudi, and Francis Bach. Infinite-dimensional sums-of-squares for optimal control. Technical Report 2110.07396, arXiv, 2021.

George S. Kimeldorf and Grace Wahba. Some results on Tchebycheffian spline functions. *J. Math. Anal. Applicat.*, 33:82–95, 1971.

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

Jean-Bernard Lasserre. The moment-SOS hierarchy and the Christoffel-Darboux kernel. Technical Report 2011.08566, arXiv, 2020.

Jean-Bernard Lasserre, Didier Henrion, Christophe Prieur, and Emmanuel Trélat. Nonlinear optimal control via occupation measures and lmi-relaxations. *SIAM Journal on Control and Optimization*, 47(4):1643–1666, 2008.

Daniel Liberzon. *Calculus of Variations and Optimal Control Theory*. Princeton University Press, 2011.

Ulysse Marteau-Ferey, Francis Bach, and Alessandro Rudi. Non-parametric models for non-negative functions. *Advances in Neural Information Processing Systems*, 33, 2020.

Ulysse Marteau-Ferey, Alessandro Rudi, and Francis Bach. Sampling from arbitrary functions via PSD models. Technical Report 2110.10527, arXiv, 2021.

Erich Novak. *Deterministic and Stochastic Error Bounds in Numerical Analysis*, volume 1349. Springer, 2006.

Pablo A. Parrilo. Semidefinite programming relaxations for semialgebraic problems. *Mathematical Programming*, 96(2):293–320, 2003.

Alessandro Rudi and Carlo Ciliberto. PSD representations for effective probability models. Technical Report 2106.16116, arXiv, 2021.

Alessandro Rudi, Ulysse Marteau-Ferey, and Francis Bach. Finding global minima via kernel approximations. Technical Report 2012.11978, arXiv, 2020.

Walter Rudin. Sums of squares of polynomials. *The American Mathematical Monthly*, 107(9):813–821, 2000.

Adrien Vacher, Boris Muzellec, Alessandro Rudi, Francis Bach, and Francois-Xavier Vialard. A dimension-free computational upper-bound for smooth optimal transport estimation. In *Conference on Learning Theory*, pages 4143–4173, 2021.

Richard Vinter. Convex duality and nonlinear optimal control. *SIAM Journal on Control and Optimization*, 31(2):518–538, 1993.

Jonathan Weed and Quentin Berthet. Estimation of smooth densities in Wasserstein distance. In *Conference on Learning Theory*, pages 3118–3119. PMLR, 2019.