Finding Global Minima via Kernel Approximations

Ulysse Marteau-Ferey

INRIA - Ecole Normale Supérieure, Paris, France

Joint work with Alessandro Rudi and Francis Bach

SMAI - June 23, 2021

Global non-convex optimization with function values

- **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 (function calls and complexity)

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

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

  - Lowest number of function calls $f(x_1), \dots, f(x_{n(\varepsilon)})$
  - Polynomial in the number of function calls n

Optimal algorithms (function calls and complexity)

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

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

  - Lowest number of function calls $f(x_1), \dots, f(x_{n(\varepsilon)})$
  - Polynomial in the number of function calls n

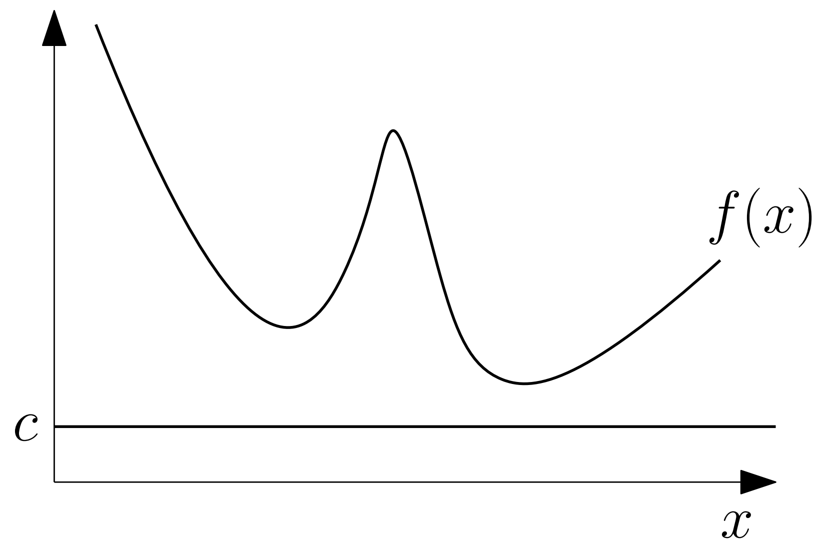
- **Optimal worst-case performance over** $\mathcal{F}$ (Novak, 2006)
  - $\mathcal{F} = m$ bounded derivatives: $n = C_{d,m} \varepsilon^{-d/m}$

- **Strategy for polynomial-time complexity in** n
  - model and optimize f **simultaeously**

Reformulation as a generic SoS problem

- **Equivalent convex problem**

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



Reformulation as a generic SOS problem

- **Equivalent convex problem**

$$\sup_{c \in \mathbb{R}} \; c \quad \text{st} \quad \forall x \in \Omega,\; f(x) - c \geqslant 0$$

- **Replace constraint** $f - c \geq 0$ **by sum of squares** $f - c = \sum_{i \in I} \lambda_i h_i^2$
  - linear model of functions $h(x) = \langle h, \phi(x) \rangle$, $\phi : \Omega \to \mathcal{H}$

$$\sup_{c \in \mathbb{R},\; \lambda \geq 0} \; c \quad \text{st} \quad \forall x \in \Omega,\; f(x) - c = \sum_{i \in I} \lambda_i \, \langle h, \phi(x) \rangle^2$$

  - **PSD problem** : writing $A = \sum_{i \in I} \lambda_i \; h_i \otimes h_i$

$$\boxed{\sup_{c \in \mathbb{R},\; A \succeq 0} \; c \quad \text{st} \quad \forall x \in \Omega,\; f(x) - c = \langle \phi(x), A\phi(x) \rangle}$$

Modeling and optimizing $f \in C^m(\Omega)$: three steps

- **Step 1 : Showing the relaxation is tight** $(1) = (2)$

$$\sup_{c \in \mathbb{R},\ A \succeq 0} \quad c \quad \text{st} \quad \forall x \in \Omega,\ \ f(x) - c = \langle \phi(x), A\phi(x) \rangle \tag{1}$$

$$\sup_{c \in \mathbb{R}} \ c \quad \text{st} \quad \forall x \in \Omega,\ \ f(x) - c \geq 0 \tag{2}$$

– SC : $\exists A_* \in S_+(\mathcal{H})$ s.t. $f(x) = f_* + \langle \phi(x), A_* \phi(x) \rangle$

Modeling and optimizing $f \in C^m(\Omega)$: three steps

- **Step 1 : Showing the relaxation is tight**
  - SC : $\exists A_* \in S_+(\mathcal{H})$ s.t. $f(x) = f_* + \langle \phi(x), A_* \phi(x) \rangle$
- **Step 2: discretizing using $n = C_{d,m} \epsilon^{-d/m}$ evaluations to have precision ϵ solving**

$$\hat{c}, \hat{A} = \underset{c \in \mathbb{R},\ A \in S_+(\mathcal{H})}{\operatorname{argmax}} \quad c - \lambda \operatorname{tr}(A) \quad \text{st} \quad f(x_i) - c = \langle \phi(x_i), A \phi(x_i) \rangle \quad (3)$$

  - guarantee that $\|\hat{c} - f_*\| \leq \epsilon$

Modeling and optimizing $f \in C^m(\Omega)$: three steps

- **Step 1 : Showing the relaxation is tight**
  - SC : $\exists A_* \in S_+(\mathcal{H})$ s.t. $f(x) = f_* + \langle \phi(x), A_* \phi(x) \rangle$
- **Step 2: discretizing using $n = C_{d,m}\epsilon^{-d/m}$ evaluations to have precision ϵ solving**

$$\hat{c}, \hat{A} = \underset{c \in \mathbb{R},\ A \in S_+(\mathcal{H})}{\operatorname{argmax}} \; c - \lambda \operatorname{tr}(A) \quad \text{st} \;\; f(x_i) - c = \langle \phi(x_i), A\phi(x_i) \rangle \quad (4)$$

- **Step 3 : showing (6) can be written as a $n \times n$ PSD program, which runs in $O(n^3)$**

$$\hat{c}, \hat{B} = \underset{c \in \mathbb{R},\ B \in S_+(\mathbf{R}^n)}{\operatorname{argmax}} \; c - \lambda \operatorname{tr}(B) \quad \text{st} \;\; f(x_i) - c = \langle \Phi_i, B\Phi_i \rangle \quad (5)$$

Modeling and optimizing $f \in C^m(\Omega)$: three steps

- **Step 1 : Showing the relaxation is tight**
  - SC : $\exists A_* \in S_+(\mathcal{H})$ s.t. $f(x) = f_* + \langle \phi(x), A_* \phi(x) \rangle$
- **Step 2: discretizing using $n = C_{d,m}\epsilon^{-d/m}$ evaluations to have precision ϵ solving**

$$\hat{c}, \hat{A} = \underset{c \in \mathbb{R},\ A \in S_+(\mathcal{H})}{\operatorname{argmax}} \ c - \lambda \operatorname{tr}(A) \quad \text{st} \quad f(x_i) - c = \langle \phi(x_i), A\phi(x_i) \rangle \quad (6)$$

- **Step 3 : showing** (6) **can be written as a $n \times n$ PSD program, which runs in $O(n^3)$**

$$\hat{c}, \hat{B} = \underset{c \in \mathbb{R},\ B \in S_+(\mathbf{R}^n)}{\operatorname{argmax}} \ c - \lambda \operatorname{tr}(B) \quad \text{st} \quad f(x_i) - c = \langle \Phi_i, B\Phi_i \rangle \quad (7)$$

- **What sould $\mathcal{H}$ be ?** a) large enough, b) finite d representation

RKHS are a natural candidate for $\mathcal{H}$

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

RKHS are a natural candidate for $\mathcal{H}$

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

- **Can represent rich spaces** Sobolev spaces $H^s(\Omega)$ with $\Omega \subset \mathbb{R}^d$, $s > d/2$

$$\langle f, g \rangle_{H^s(\Omega)} = \sum_{|\alpha| \leq s} \int_{\Omega} \partial^\alpha f \; \partial^\alpha g$$

The kernel k can be computed explicitly with Bessel functions

RKHS are a natural candidate for $\mathcal{H}$

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

- **Can represent rich spaces** Sobolev spaces $H^s(\Omega)$ with $\Omega \subset \mathbb{R}^d$, $s > d/2$

- **Made for sample-based approaches : representer theorem**
  - Problem $\min_{g \in \mathcal{H}} L(g(x_1), ..., g(x_n)) + \frac{\lambda}{2} \|g\|_{\mathcal{H}}^2, \; \lambda \geq 0$
  - Finite dimensional representer theorem in $\mathbb{R}^n$:

$$g_{\text{opt}}(x) = \sum_{i=1}^{n} \alpha_i \, k(x_i, x) \quad \Longrightarrow \quad \text{becomes problem in } \alpha$$

Step 1 : showing that f is SoS

Theorem: Assume Ω is bounded, $f \in C^m(\Omega)$ has isolated strict-second order minima in $\overset{\circ}{\Omega}$ and is greater than $\delta > 0$ near the boundary $\partial\Omega$.

For any $d/2 < s \leq m-2$, there exists $h_1, ..., h_N \in H^s(\Omega)$ such that

$$\forall x \in \Omega, \;\; f(x) = f_* + \sum_{i=1}^{N} h_i^2(x)$$

$$= f_* + \langle \phi(x), A_* \phi(x) \rangle_{H^s(\Omega)}$$

where $A_* = \sum h_i \otimes h_i$

Step 1 : showing that $f - f_*$ is SoS (proof sketch 1)

- **Assumption**: Assume Ω is bounded, $f \in C^m(\Omega)$ has isolated strict-second order minima in $\overset{\circ}{\Omega}$ and is greater than $\delta > 0$ near the boundary $\partial\Omega$.

- **From local to global** If $f - f_*$ is SoS locally, then it is SoS globally compactness argument + gluing with partition of unity of the form

$$1 = \sum_{i=1}^{N} \chi_i^2$$

- If $f(x_0) - f_* > 0$, then $f(x) - f_* > \delta$ locally and hence $\sqrt{f - f_*} \in C^m(B(x_0, r_0)) \subset H^s(B(x_0, r_0))$

Step 1 : showing that $f - f_*$ is SoS (proof sketch 2)

- If $f(x_0) - f_* = 0$, then locally (strict minimum assumption)

$$f(x) - f_* = \frac{1}{2}(x - x_0)^\top \underbrace{\left(\int_0^1 (1-t)\nabla^2 f(x_0 + t(x - x_0))dt \right)}_{R(x) \in H^s(B(x_0, r_0)) \succ \delta I} (x - x_0)$$

- $\sqrt{R(x)} \in H^s(B(x_0, r_0))$

- $h(x) = \sqrt{R(x)}(x - x_0) \in H^s(B(x_0, r_0)),\ f - f_* = \sum h_i^2$

Step 2 : discretizing using random samples

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

$$\hat{c}, \hat{A} = \underset{c \in \mathbb{R},\ A \succcurlyeq 0}{\operatorname{argmax}} \, c - \lambda \, \mathrm{tr}(A) \quad \text{st} \quad f(x_i) = c + \langle \phi(x_i), A\phi(x_i) \rangle$$

- **Theorem** (Rudi, Marteau-Ferey, and Bach, 2020) Up to logarithmic terms : if $n = C_{d,m,\Omega}\, \varepsilon^{-d/(m-d/2-3)}$ and the samples $(x_1, ..., x_n)$ are taken randomly from Ω, and if $\lambda = \varepsilon$, then it holds with probability at least $1 - \delta$:

$$|\hat{c} - f_*| \leq \varepsilon \ \ \mathrm{tr}(A_*) \ \ \log \tfrac{1}{\delta}$$

- **Optimal rates** : $n = C_{d,m,\Omega} \epsilon^{-d/(m-d/2)}$

Step 2 : discretizing using random samples (proof ideas)

- **Scattered data inequality** If $(x_1, ..., x_n)$ δ coverage of Ω, then

$$\begin{aligned} |f(x) - \hat{c} - \langle \phi(x), \hat{A}\phi(x) \rangle| &\leq \|f - c - g_A\|_{C^{m-3-d/2}} \, \delta^{m-3-d/2} \\ &\leq (\mathrm{tr}(A_*) + \mathrm{tr}(\hat{A})) \, \delta^{m-3-d/2} \end{aligned}$$

Conclusion : $\hat{c} - f_* \leq (\mathrm{tr}(A_*) + \mathrm{tr}(\hat{A})) \, \delta^{m-3-d/2}$

Step 2 : discretizing using random samples (proof ideas)

- **Scattered data inequality** If $(x_1, ..., x_n)$ δ coverage of Ω, then

$$\begin{aligned} |f(x) - \hat{c} - \langle \phi(x), \hat{A}\phi(x) \rangle| &\leq \|f - c - g_A\|_{C^{m-3-d/2}} \, \delta^{m-3-d/2} \\ &\leq (\mathrm{tr}(A_*) + \mathrm{tr}(\hat{A})) \, \delta^{m-3-d/2} \end{aligned}$$

  Conclusion : $\hat{c} - f_* \leq (\mathrm{tr}(A_*) + \mathrm{tr}(\hat{A})) \, \delta^{m-3-d/2}$

- If $(x_1, ..., x_n)$ sampled randomly, up to log factors, it is a $\delta = n^{-1/d}$ coverage of Ω

  Conclusion : $\hat{c} - f_* \leq (\mathrm{tr}(A_*) + \mathrm{tr}(\hat{A})) \, n^{-\frac{m-3-d/2}{d}}$

- **Bound for the regularizing term** bound $\mathrm{tr}(\hat{A})$ in terms of $\mathrm{tr}(A_*)$

Step 3 : Finite dimensional formulation

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

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

- **Finite dimensional problem** Restriction to $\mathcal{H}_n = vect(\phi(x_i))$:

  $A \in S_+(\mathcal{H}) \longrightarrow A \in S_+(\mathcal{H}_n)$

Step 3 : Finite dimensional formulation

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

$$\sup_{c\in\mathbb{R},\ A\succcurlyeq 0} c - \lambda\, \mathrm{tr}(A) \ \text{ s.t. } \ \forall i \in \{1,\ldots,n\},\ f(x_i) = c + \langle \phi(x_i), A\phi(x_i)\rangle$$

- **Finite dimensional problem** Restriction to $\mathcal{H}_n = vect(\phi(x_i))$:

  $A \in S_+(\mathcal{H}) \longrightarrow A \in S_+(\mathcal{H}_n)$

- **Finite-dimensional formulation** : Representer theorem for RKHS SoS (Marteau-Ferey, Bach, and Rudi (2020))

  SDP of dimension n :

$$\sup_{c\in\mathbb{R},\ B\succcurlyeq 0, B\in\mathbb{R}^{n\times n}} c - \lambda\, \mathrm{tr}(B) \ \text{ st } \forall i \in \{1,\ldots,n\},\ f(x_i) = c + \Phi_i^\top B \Phi_i$$

- **Solvable in polynomial time** with precision ϵ in $O(n^{3.5} \log \frac{1}{\epsilon})$

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**

   – Set $f_j = f(x_j)$, $\forall j \in \{1, \dots, n\}$
   – Compute $K_{ij} = k(x_i, x_j)$ for k Sobolev kernel of smoothness s
   – Set $\Phi_j \in \mathbb{R}^n$ computed using a cholesky decomposition of K $\forall j \in \{1, \dots, n\}$.

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

- **Output:** c proxy for f_*

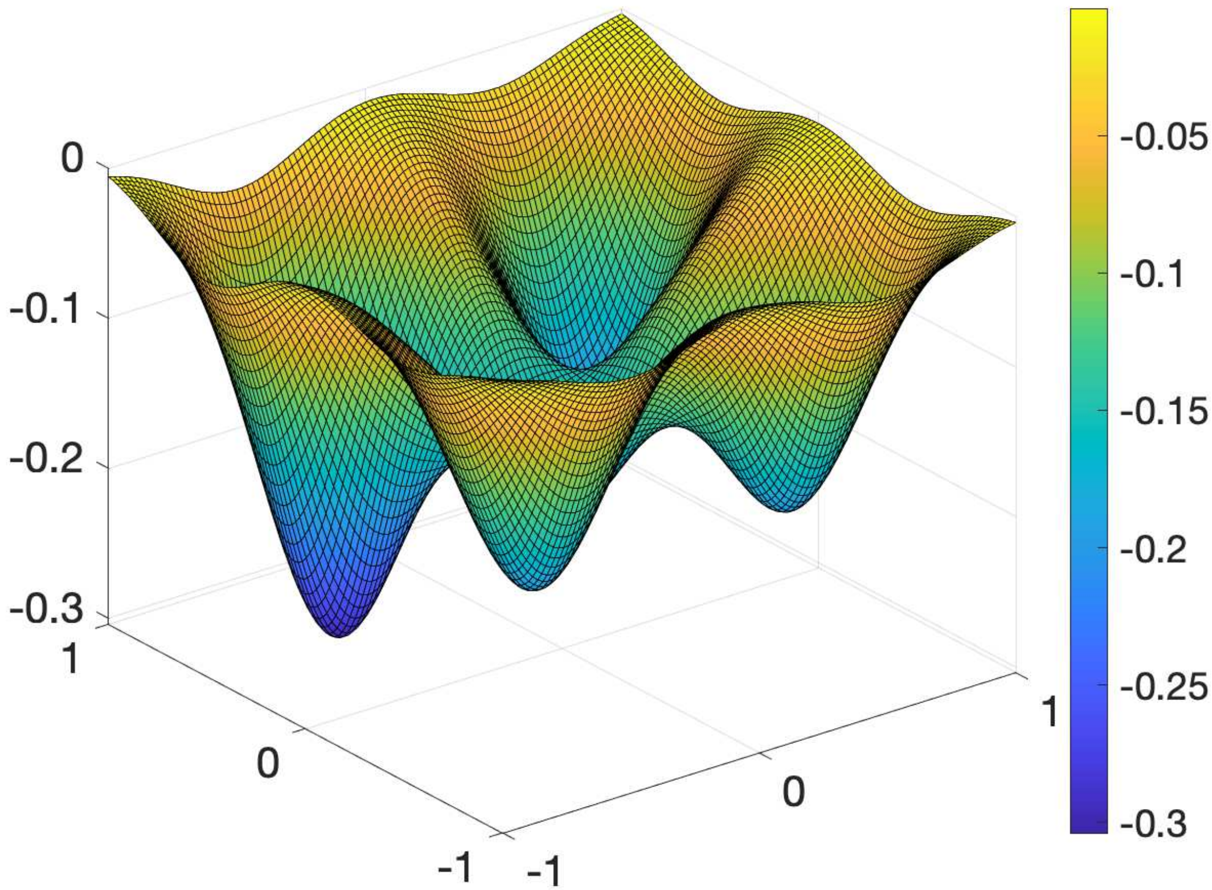
- One can extend the algorithm in order to compute a proxy of the minimizer

Main properties of the model

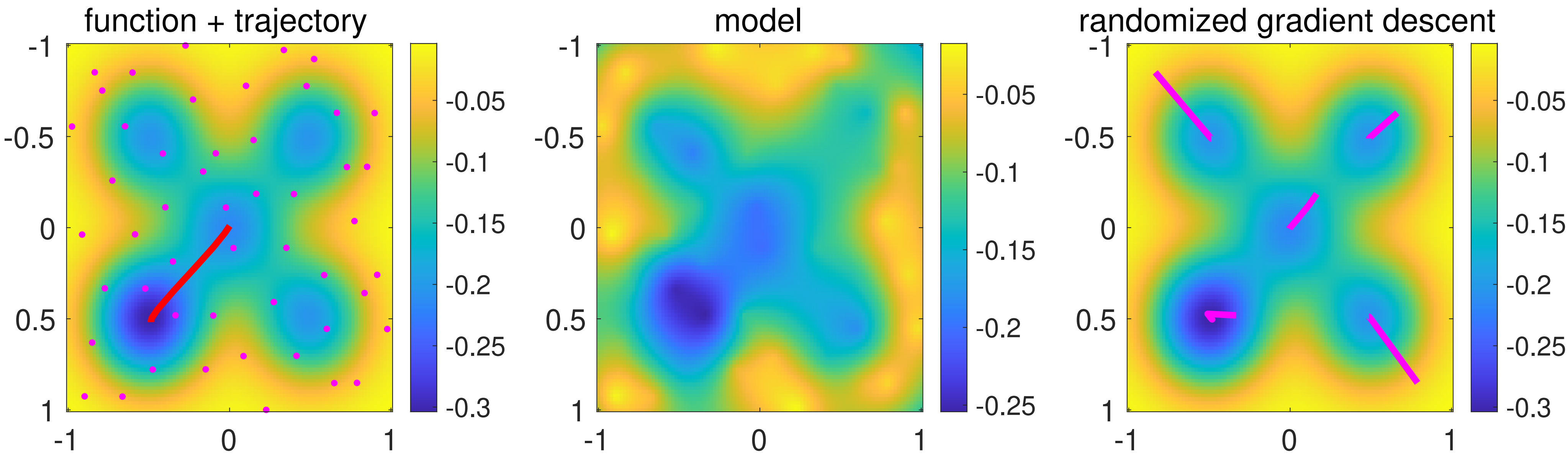
- "Always" possible to write a non-negative function as a RKHS SoS
- Bounds on the number of samples needed for a given precision
- Finite dimensional SDP with bounded complexity $O(n^{3.5} \log \frac{1}{\epsilon})$
- Breaks the curse of dimensionality in term of sample numbers (needs $\epsilon^{-d/m}$ samples) for smooth enough functions (but not in the constants)
- For the moment, **no certificate bound on the result of the algorithm**

Illustration

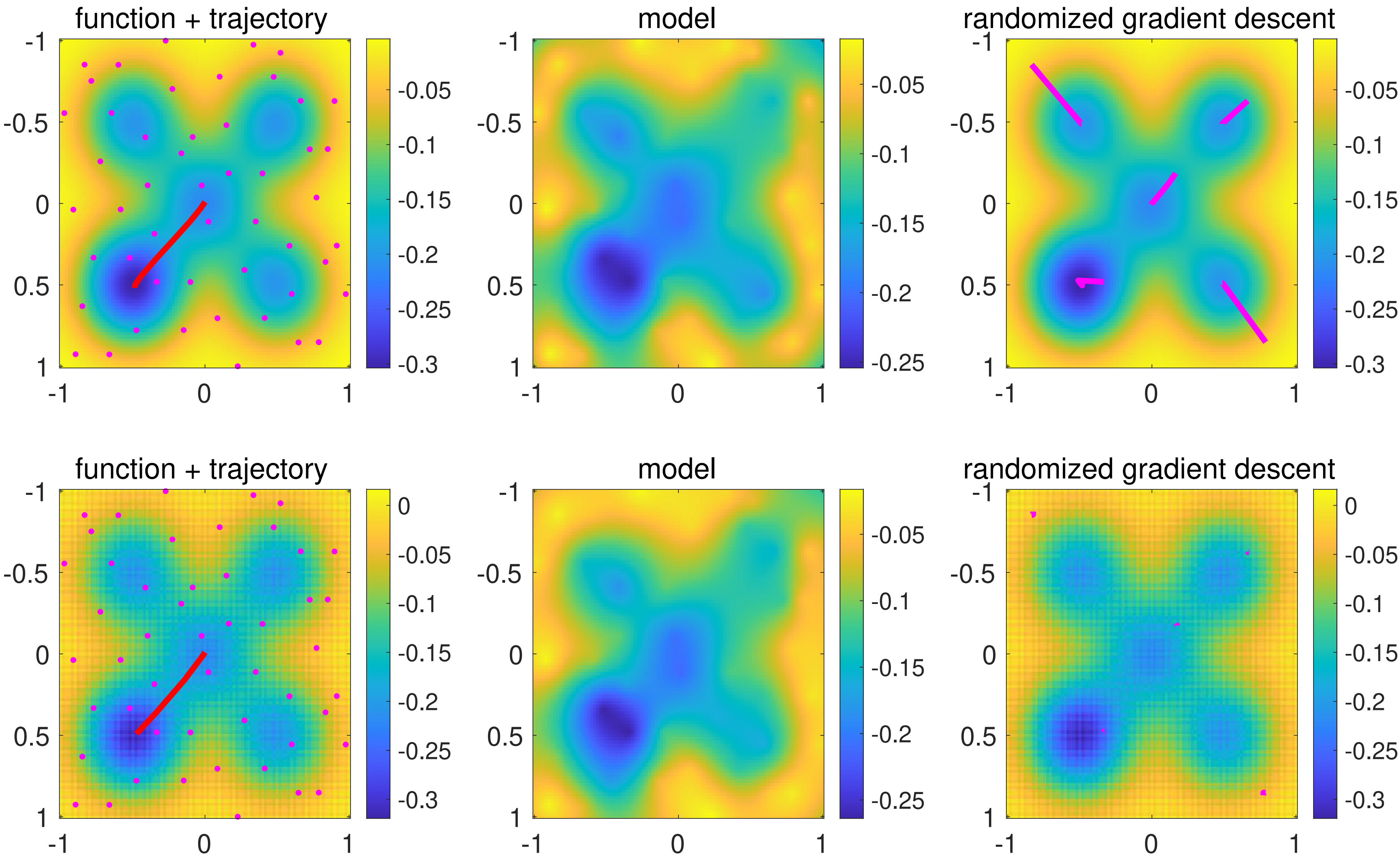
- **Minimization of two-dimensional function**



Illustration

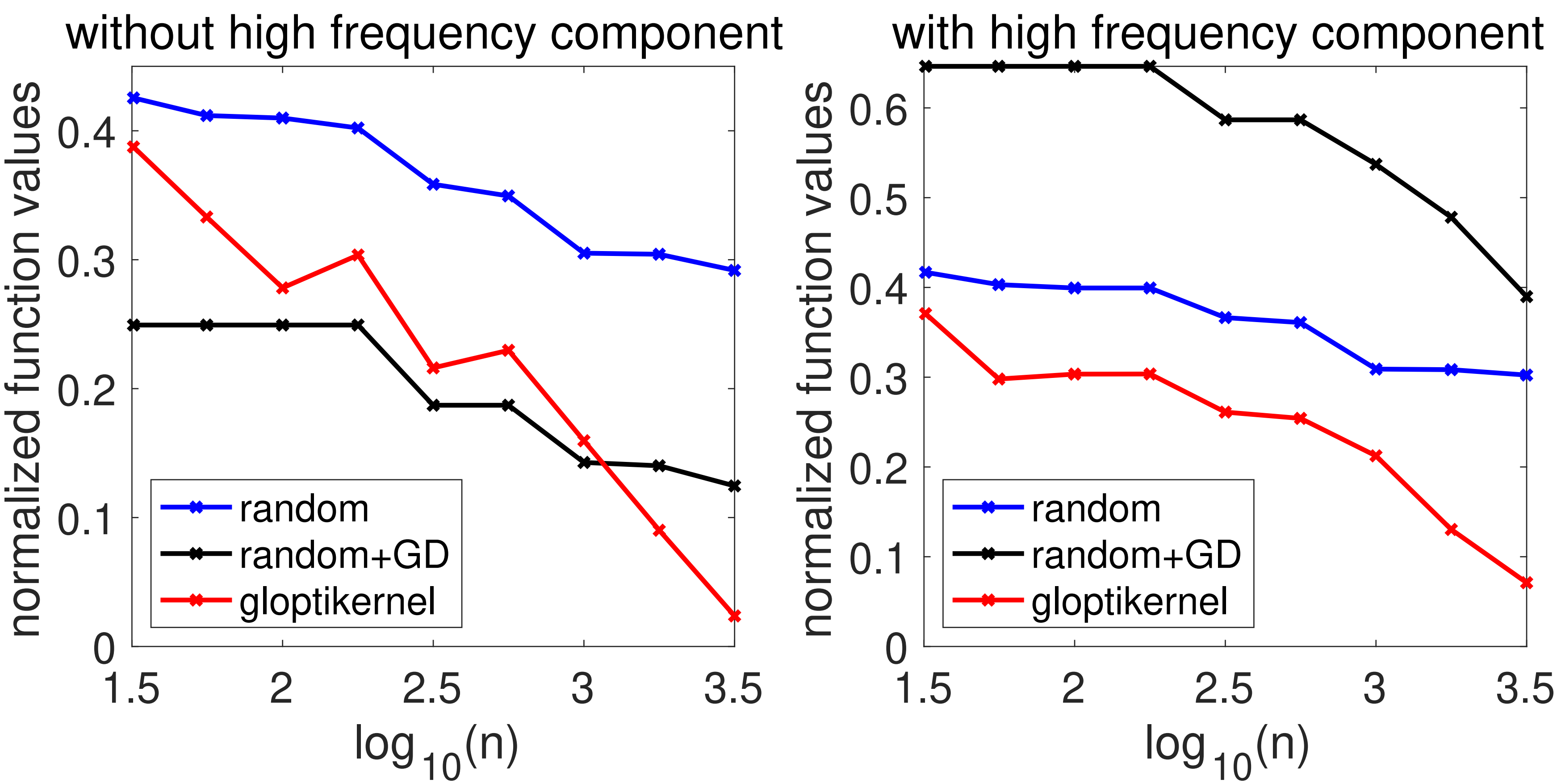


Illustration



Illustration

- **Minimization of eight-dimensional function**



Extension

- **Constrained optimization problem**

$$\inf_{x \in \mathbb{R}^d} f(x) \quad \text{such that} \quad \forall x \in \Omega, \ g(x) \geqslant 0$$

Extension

- **Constrained optimization problem**

$$\inf_{x \in \mathbb{R}^d} f(x) \quad \text{such that} \quad \forall x \in \Omega,\ g(x) \geqslant 0$$

- **Sums-of-squares reformulation**

$$\sup_{c \in \mathbb{R},\ A \succcurlyeq 0,\ B \succcurlyeq 0} \quad c$$

$$\text{such that} \quad \forall x \in \Omega,\ f(x) = c + \langle \phi(x), A\phi(x) \rangle + g(x) \langle \phi(x), B\phi(x) \rangle$$

  – Extension of results on polynomials (Lasserre, 2001)

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 extensions**
  - Efficient algorithms below $O(n^3)$ complexity
  - Adaptive choice of sampling points
  - Other infinite-dimensional convex optimization problems

Conclusion

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

- **Further extensions**
  - Efficient algorithms below $O(n^3)$ complexity
  - Adaptive choice of sampling points
  - Other infinite-dimensional convex optimization problems

- **See** `arxiv.org/abs/2012.11978` and `francisbach.com/` for interesting blog posts !

References

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

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

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

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