Advanced models for function representation

To improve adaptivity in applied math algorithms

Alessandro Rudi
INRIA and École Normale Supérieure, Paris

04 April 2024 — MASCOT-NUM, Hyeres

Collaborators

- Ulysse Marteau-Ferey, Eloise Berthier, Gaspard Beugnot, Theophile Cantelobre, Adrien Vacher
- Boris Muzellec, Pierre-Cyril Aubin, Blake Woodworth
- Francis Bach, Carlo Ciliberto, Justin Carpentier, FX Vialard

ERC starting grant
“REAL” 2021-2026

Dream algorithm

in applied math (in a world of big data…)

- Certified by theory
- Adaptive to the regularity of the problem
  (few resources, when the problem is easier)

State of the art

Do we have certified *and* adaptive algorithms?

Yes (in important cases)

- Solution of PDEs (meshless methods)
- Quadrature and integration
- Interpolation and approximation…

open problem

- (Non-convex) Optimization
- Optimal Transport
- Optimal Control…

How to obtain systematically adaptivity?

Example: Optimal Transport

Computation of Wasserstein distance from samples

$$\mathrm{OT}(\mu,\nu) = \sup_{u,v \in C(\mathbb{R}^d)} \int u(x)d\mu(x) + \int v(y)d\nu(y)$$

$$c(x,y) \geq u(x) + v(y), \quad \forall (x,y) \in X \times Y,$$

Assumption 1 (m-times differentiable densities) *Let $m, d \in \mathbb{N}$. Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$.*

a) μ, ν have densities. Their supports, resp. $X, Y \subset \mathbb{R}^d$ are convex, bounded and open;

b) the densities are finite, bounded away from zero, with Lipschitz derivatives up to order m.

Caffarelli et al. 1992

Example: Optimal Transport

Computation of Wasserstein distance from samples

$$\mathrm{OT}(\mu,\nu) = \sup_{u,v \in C(\mathbb{R}^d)} \int u(x)d\mu(x) + \int v(y)d\nu(y)$$

$$c(x,y) \geq u(x) + v(y), \quad \forall (x,y) \in X \times Y,$$

State of the art:

Samples $O(\varepsilon^{-\frac{2m-1+d/2}{m}})$

Complexity $O(\varepsilon^{-2d})$

Hutter, Rigollet 2019; Niles-Weed, Rigollet 2019; Weed, Berthet 2020;

Function representation in applied math

Linear models

Crucial tool to crack some problems of applied math

$$f_\theta(x) \;=\; \sum_{i=1}^{n} \theta_i \, g(x, x_i)$$

- Solution of PDEs (meshless methods)
- Finite element methods
- Quadrature and integration problems
- Interpolation and approximation
- machine learning ….

Good properties of linear models

for problems in applied mathematics

$$f_\theta(x) \;=\; \sum_{i=1}^{n} \theta_i \, g(x, x_i)$$

$g(x, x')$ kernel function, e.g. $e^{-\eta \|x - x'\|^2}$

$x_1, \ldots, x_n$ given, e.g. sampled uniformly at random

$\theta_i \in \mathbb{R}$

1) Linearity in the parameters

- $L(f)$ convex $\rightarrow$ $\min_{\theta \in \mathbb{R}^n} L(f_\theta)$ convex
- $\int f_\theta(x) dx$ in closed form for nice $g(x, x_i)$

Good properties of linear models

for problems in applied mathematics

$$f_\theta(x) \;=\; \sum_{i=1}^{n} \theta_i \, g(x, x_i)$$

$g(x, x')$ kernel function, e.g. $e^{-\eta \|x - x'\|^2}$

$x_1, \ldots, x_n$ given, e.g. sampled uniformly at random

$\theta_i \in \mathbb{R}$

2) Concise approximation

Thm. There exist $x_1, \ldots, x_n$ such that, when

$$n = O(\varepsilon^{-\frac{d}{m}})$$

For any $u \in C^m([-1,1]^d)$, there exists $\theta \in \mathbb{R}^n$ such that

$$\|u - f_\theta\|_{L^\infty} \leq C_u \varepsilon$$

Scattered Data Approximation, Wendland 2004

Few points suffice to approximate smooth functions!

However, sometimes, linear models are not enough…

Sometimes linear models are not enough…

Many interesting problems need conic structure

- Probability representation, bayesian inference
- (Non-convex) Optimization
- Uncertainty quantification
- Optimal Transport
- Optimal Control
- Reinforcement Learning
- … .

$$p(x) \geq 0, \quad \int p(x) = 1$$

linear models, and other important models miss crucial properties...

… they need more structure: non-negativity

E.g. probability representation

$$p(x) \geq 0, \quad \int p(x) = 1$$

Linear models

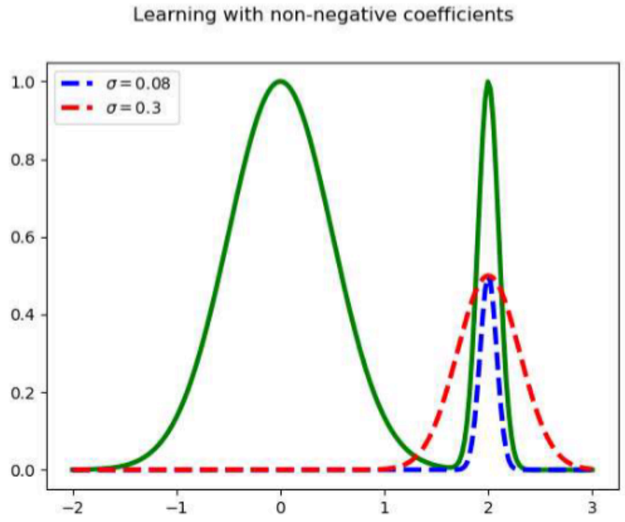
- Concise approximation
- Linearity in the parameters
- Non-negativity

Linear models cannot be used for probability representation

State of the art of models are not enough

For non-negative functions

- **Ridge models** e^{-f_θ}
  - Concise approximation ✓
  - Linearity in the parameters ✗
- **Mixture models** $\sum_{i=1}^{n} \theta_i g(x, x_i) \quad \theta_i \geq 0, g(\cdot, x_i) \geq 0$
  - Concise approximation ✗ they need $n = O(\varepsilon^{-d/2})$
  - Linearity in the parameters ✓



Tsybakov, *Introduction to Nonparametric Estimation*, Springer Series in Statistics, 2008
Sriperumbudur et al. *Density estimation in infinite dimensional exponential families*. *JMLR 2017*.

A new model for non-negative functions

New model: PSD representation

It has all the key properties

$$f_A(x) = \sum_{i,j=1}^{n} A_{i,j} g(x, x_i) g(x, x_j) \qquad A \in \mathbb{R}^{n \times n}, A \succeq 0$$

- $f_A(x) \geq 0$ for any x, by construction (easy generalisation to cones)
- Indeed $f_A(x) = \Phi(x)^\top A \Phi(x) \geq 0, \quad \forall x \in X, \quad \Phi(x) = (g(x, x_1), \dots, g(x, x_n))$
- More generally $\Phi : X \to \mathcal{H}$ feature map of some reproducing kernel

Marteau-Ferey, Bach, Rudi. "Non-parametric models for non-negative functions." NeurIPS (2020)

New model: PSD representation

It has all the key properties

$$f_A(x) = \sum_{i,j=1}^{n} A_{i,j} g(x, x_i) g(x, x_j) \qquad A \in \mathbb{R}^{n \times n}, A \succeq 0$$

- $f_A(x) \geq 0$ for any x, by construction (easy generalisation to cones)
- f_A is linear in A, then
  - Preserves convexity: $\min_{A \succeq 0} L(f_A)$ convex, when $L(f)$ is convex
  - Integrals and linear operators in closed form for nice $g(x, x_i)$

Marteau-Ferey, Bach, Rudi. "Non-parametric models for non-negative functions." NeurIPS (2020)

New model: PSD representation

It has all the key properties

$$f_A(x) = \sum_{i,j=1}^{n} A_{i,j} g(x, x_i) g(x, x_j) \qquad A \in \mathbb{R}^{n \times n}, A \succeq 0$$

- $f_A(x) \geq 0$ for any x, by construction (easy generalisation to cones)
- f_A is linear in A, then
  - Preserves convexity: $\min_{A \succeq 0} L(f_A)$ convex, when $L(f)$ is convex
  - Integrals and linear operators in closed form for nice $g(x, x_i)$
- Concise approximator: small n suffices for approximate well smooth functions!

Marteau-Ferey, Bach, Rudi. "Non-parametric models for non-negative functions." NeurIPS (2020)

PSD models are concise approximators (1)

Marteau-Ferey, Bach, R. 2020

Starter: PSD models concisely approximate the family of exponential models

- Goal: approximate $f = \exp(V(x)), \quad V \in C^m([-1,1]^d)$
- Define $h \in C^m([-1,1]^d)$ s.t. $f = h^2, \quad h = \exp(V/2)$
- Find the linear model f_a, that approximates h

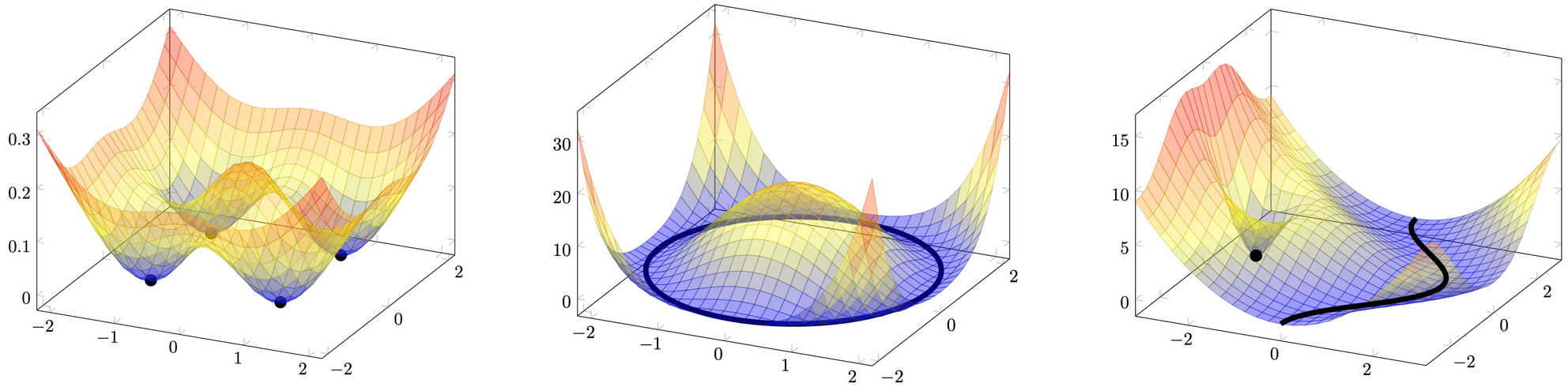
$$sup_{x\in[-1,1]^d} |f_a(x) - h(x)| \leq \epsilon, \quad f_a(x) = \sum_{i=1}^{n} a_i g(x, x_i) \qquad n \approx \epsilon^{-d/m}$$

- Build the PSD model $f_A = f_a^2, \quad A = aa^\top$
- Then

$$\sup_{x\in[-1,1]^d} |f_A(x) - f(x)| \leq C\epsilon, \quad f_A(x) = \sum_{ij=1}^{n} A_{ij} g(x, x_i) g(x, x_j) \qquad n \approx \epsilon^{-d/m}$$

Marteau-Ferey, Bach, Rudi. "Non-parametric models for non-negative functions." NeurIPS (2020)

PSD models are concise approximators (2)



2) Concise approximation

Thm. When $u \geq 0, u \in C^m([-1,1]^d)$. The zeros are a manifold. Locally the Hessian of u is strictly positive in the direction orthogonal to this manifold.
Then there exists $A \in \mathbb{R}^{n \times n}, A \succeq 0$ s.t. $\sup_{x \in [-1,1]^d} |f_A(x) - u(x)| \leq C\epsilon,$ $n \approx \epsilon^{-d/m}$

Rudi, Ciliberto. "PSD Representations for Effective Probability Models." NeurIPS (2021).

Back to Optimal Transport

Vacher, Muzellec, Rudi, Bach, Vialard. ***A Dimension-free Smooth Optimal Transport Estimation. COLT 2021***

Application: Optimal Transport

Concise approximation of the inequality

$$\mathrm{OT}(\mu,\nu) = \sup_{u,v \in C(\mathbb{R}^d)} \int u(x)d\mu(x) + \int v(y)d\nu(y)$$

$$c(x,y) \geq u(x) + v(y), \quad \forall (x,y) \in X \times Y,$$

THEOREM 1.4.
Let $n, m, s, d \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$ and bounded. Let $\mathcal{H}(\Omega) = W_2^s(\Omega)$ with $s > m + d/2$. Let

$$n = O(\varepsilon^{-d/m}).$$

There exists $x_1, \ldots, x_n \in \Omega$ such that the following holds.
For any $h \in W_\infty^m(\Omega)$, $A \in \mathcal{L}(\mathcal{H}(\Omega))$ and $A \succeq 0$ we have

$$h(x_i) = f_A(x_i),\ i = 1, \ldots, n \quad \Rightarrow \quad h(x) \geq -(\|h\|_m + \mathrm{tr}(A))\,\varepsilon \quad \forall x \in \Omega.$$

Rudi, Marteau-Ferey, Bach. *Finding Global Minima via Kernel Approximations.* Mathematical Programming. (to appear 2024)

Application: Optimal Transport

Let's transform the inequality into an equality

$$\mathrm{OT}(\mu,\nu) = \sup_{u,v \in C(\mathbb{R}^d)} \int u(x)d\mu(x) + \int v(y)d\nu(y)$$

$$f \geq 0 \qquad c(x,y) - u(x) - v(y) = f(x,y) \quad \forall\ (x,y) \in X \times Y$$

Application: Optimal Transport

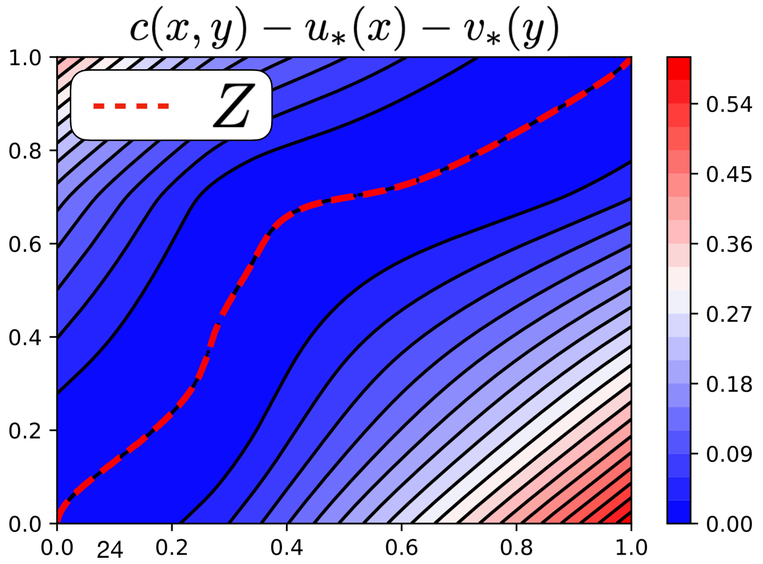
Let's use Assumption 1

$$\mathrm{OT}(\mu,\nu) = \sup_{\substack{u,v \in C^m \\ f \geq 0, f \in C^m}} \int u(x)d\mu(x) + \int v(y)d\nu(y)$$
$$c(x,y) - u(x) - v(y) = f(x,y) \quad \forall\ (x,y) \in X \times Y$$

Caffarelli implies

$$u \in C^m(X), v \in C^m(Y)$$

and regular zeros



Application: Optimal Transport

Subsample the inequality

$$\mathrm{OT}(\mu,\nu) = \sup_{\substack{u,v \in C^m \\ f \geq 0, f \in C^m}} \int u(x)d\mu(x) + \int v(y)d\nu(y)$$
$$c(x,y) - u(x) - v(y) = f(x,y) \quad \forall\ (x,y) \in X \times Y$$

Now we can use the power of Approximation theory

$$h(x_i,y_i) = 0, \quad \forall i = 1,\ldots,n \quad \Rightarrow \quad |h(x,y)| \leq \varepsilon, \quad \forall (x,y) \in X \times Y$$

when $n = O(\ \|h\|_{C^m} \varepsilon^{-d/m}\ )$

we apply this result to $h(x,y) := c(x,y) - u(x) - v(y) - f(x,y)$

Application: Optimal Transport

Subsample the inequality

$$\widehat{OT} = \sup_{\substack{u,v \in C^m \\ f \geq 0, f \in C^m}} \int u(x)d\mu(x) + \int v(y)d\nu(y)$$
$$c(x_i, y_i) - u(x_i) - v(y_i) - f(x_i, y_i) = 0, \quad i = 1, \ldots, n$$

Now we can use the power of Approximation theory

$$h(x_i, y_i) = 0, \quad \forall i = 1, \ldots, n \quad \Rightarrow \quad |h(x, y)| \leq \varepsilon, \quad \forall (x, y) \in X \times Y$$

when $n = O(\ \|h\|_{C^m} \varepsilon^{-d/m}\ )$

we apply this result to $h(x, y) := c(x, y) - u(x) - v(y) - f(x, y)$

Application: Optimal Transport

Represent f with PSD models

$$\widehat{OT} = \sup_{\substack{u,v \in C^m \\ f \geq 0, f \in C^m}} \int u(x)d\mu(x) + \int v(y)d\nu(y)$$
$$c(x_i, y_i) - u(x_i) - v(y_i) - f(x_i, y_i) = 0, \quad i = 1, \ldots, n$$

we need a model
- for non-negative functions
- that preserves convexity
- that is concise approximator

use PSD models (and Thm. 1.4)

Application: Optimal Transport

Represent f with PSD models

$$\widehat{OT} = \sup_{\substack{u,v \in C^m \\ f \in PSD(\varepsilon)}} \int u(x)d\mu(x) + \int v(y)d\nu(y)$$
$$c(x_i, y_i) - u(x_i) - v(y_i) - f(x_i, y_i) = 0, \quad i = 1, \ldots, n$$

we need a model
- for non-negative functions
- that preserves convexity
- that is concise approximator

use PSD models (and Thm. 1.4)

Application: Optimal Transport

Approximate the rest

$$\widehat{OT} = \sup_{\substack{u,v \in \mathcal{H} \\ f \in PSD(\varepsilon)}} \frac{1}{M}\sum_{j=1} u(\tilde{x}_j) + \frac{1}{M}\sum_{j=1} v(\tilde{y}_j)$$
$$c(x_i, y_i) - u(x_i) - v(y_i) - f(x_i, y_i) = 0, \quad i = 1, \ldots, n$$

Easy approximation of the rest:

- u, v with linear models
- the integrals with Monte Carlo

Application: Optimal Transport

Approximate the rest

$$\widehat{OT} = \sup_{\substack{u,v \in \mathcal{H} \\ f \in PSD(\varepsilon)}} \frac{1}{M}\sum_{j=1} u(\tilde{x}_j) + \frac{1}{M}\sum_{j=1} v(\tilde{y}_j)$$
$$c(x_i, y_i) - u(x_i) - v(y_i) - f(x_i, y_i) = 0, \quad i = 1, \ldots, n$$

-This is a Semidefinite Programming Problem

-Solvable with damped Newton method in $O(n^{3.5}\log(1/\varepsilon) + nM + M^2)$

Example: Optimal Transport

Sketch of the final theorem

THEOREM 1.1. *Let $\delta, \varepsilon \in (0, 1]$ and $m > d$. Under Assumption 1, when*

$$M = O(\varepsilon^{-\frac{2m-1+d/2}{2m}} \log(1/\delta)), \qquad n = O(\varepsilon^{-\frac{2d}{m-d}}),$$

then with probability at least $1 - \delta$

$$|\widehat{OT} - OT(\mu, \nu)| \leq O(\varepsilon).$$

Moreover, the computation of $\widehat{OT}$ requires

$$\tilde{O}(\varepsilon^{-\frac{7d}{m-d}} + \varepsilon^{-4}).$$

Vacher, Muzellec, Rudi, Bach, Vialard. *A Dimension-free Smooth Optimal Transport Estimation. COLT 2021*

Application: Optimal Transport

Computation of Wasserstein distance between densities

$$\mathrm{OT}(\mu,\nu) = \sup_{u,v \in C(\mathbb{R}^d)} \int u(x)d\mu(x) + \int v(y)d\nu(y)$$

$$c(x,y) \geq u(x) + v(y), \quad \forall (x,y) \in X \times Y,$$

State of the art:

Samples $O(\varepsilon^{-\frac{2m-1+d/2}{m}})$

Complexity $O(\varepsilon^{-2d})$

Algorithm with PSD model:

$O(\varepsilon^{-\frac{2m-1+d/2}{2m}})$

$\tilde{O}(\varepsilon^{-\frac{7d}{m-d}} + \varepsilon^{-4})$

Hutter, Rigollet 2019; Niles-Weed, Rigollet 2019; Weed, Berthet 2020;

Vacher, Muzellec, Rudi, Bach, Vialard, 2021.

Many interesting problems have non-negativity constr.

- Probability representation, bayesian inference
- (Non-convex) Optimization
- Uncertainty quantification
- Optimal Transport
- Optimal Control
- Reinforcement Learning
- … .

$$p(x) \geq 0, \quad \int p(x) = 1$$

linear models, and other important models miss crucial properties...

Infinite-dim. convex opt. with dense conic constraints

A framework for many problems in applied math

$$\inf_{u \in V} L(u)$$
$$G(x,u) = 0, \quad \forall x \in \Omega_1$$
$$H(x,u) \geq 0, \quad \forall x \in \Omega_2$$

Possibly infinite-dim. (Typically a function space)

Dense set of conic constraints (H affine in u)

- V Separable Banach space
- $L : V \to \mathbb{R}$ Convex functional to be minimised
- Ω_1, Ω_2 subsets of a metric space, e.g. $\mathbb{R}^d$
- $G : \Omega_1 \times V \to \mathbb{R}$ function, affine in V
- $H : \Omega_2 \times V \to \mathbb{R}$ function, affine in V

Ekeland, Temam. *Convex analysis and variational problems*. Society for Industrial and Applied Mathematics, 1976.
Ekeland, Turnbull. *Infinite-dimensional optimization and convexity*. University of Chicago Press, 1983.

Same approach for other problems

Transform problems with dense inequalities into PSD models + subsampling

Probability representation & inference

- **Rudi, Ciliberto. *PSD Representations for Effective Probability Models.* NeurIPS 2021.**
- Marteau-Ferey, Bach, Rudi. *Sampling from Arbitrary Functions via PSD Models.* **AISTATS 2021**
- Marteau-Ferey, Bach, Rudi. *Non-parametric models for non-negative functions.* **NeurIPS 2020**.

Non-convex Optimization

- **Rudi, Marteau-Ferey, Bach. *Finding Global Minima via Kernel Approximations.* Mathematical Programming. (to appear 2024).**
- Marteau-Ferey, Bach, Rudi. *Second-order conditions to decompose smooth functions as SoS.* **SIAM Journal on Optimization 2024.**
- Bach, Rudi. *Exponential Convergence of SoS hierarchies for trigonometric polynomials.* **SIAM Journal on Optimization 2023.**
- Beugnot, Mairal, Rudi. *Gloptinets: scalable Non-convex optimization with certificates*. **NeurIPS 2023.**

Learning & Simulating Stochastic Differential Equations

- **Bonalli, Rudi. *Non-parametric Learning of SDEs with fast rates.* Foundations of Computational Mathematics (subm. 2024).**
- Raj, Simsekli, Rudi. *Efficient sampling of SDEs with Positive Semidefinite models.* **NeurIPS 2023.**

Optimal Transport

- Muzellec, Vacher, Bach, Vialard, Rudi. *Near-optimal estimation of OT maps via k-SoS.* **SIAM J. on Math. and Data Science (to app. 2024)**
- **Vacher, Muzellec, Rudi, Bach, Vialard. *A Dimension-free Smooth Optimal Transport Estimation.* COLT 2021**

Optimal Control

- Bertier, Rudi, Bach, Carpentier. Optimal Control via PSD models. **CDC 2022**

Semi-infinite programming

- **Aubin, Rudi. Approximation of Optimization Problems via kernel SoS. Optimization 2024.**

Conclusion & open directions

- We propose an extended collocation method to approximate infinite-dimensional problems with dense conic constraints
- It guarantees 1. Adaptivity to the regularity of the instance under clear geometric conditions
  2. Automatic finite-dimensional formulation as SDP
  3. For regular instances its cost overcomes the curse of dimensionality in the rate
- It allows to derive adaptive algorithms for some relevant problems in applied math

Open directions:

- Can we achieve faster formulations, beyond SDP?
- Often hidden constants are exp(d), can we characterise easier subclasses of problems?
- Can we obtain certificates a posteriori for the obtained solution?

$$\max_{\substack{u\in\mathcal{H}_X, v\in\mathcal{H}_Y,\\ A\in\mathbb{S}_+(\mathcal{H}_{XY})}} \quad \langle u,\, \hat{w}_\mu\rangle_{\mathcal{H}_X} + \langle v,\, \hat{w}_\nu\rangle_{\mathcal{H}_Y} - \lambda_1 \mathrm{Tr}(A) - \lambda_2(\|u\|^2_{\mathcal{H}_X} + \|v\|^2_{\mathcal{H}_Y})$$

$$\text{subject to} \quad \forall j \in [\ell], \quad c(\tilde{x}_j, \tilde{y}_j) - u(\tilde{x}_j) - v(\tilde{y}_j) = \langle \phi(\tilde{x}_j, \tilde{y}_j),\, A\phi(\tilde{x}_j, \tilde{y}_j)\rangle_{\mathcal{H}_{XY}}.$$

Comparison with other methods

Global optimisation with SoS polynomials

(univariate trigonometric: **Nesterov, 2000**; multivariate **Lasserre, 2000**)

$$\hat{c} = \max_{c \in \mathbb{R}, A \succeq 0} c \quad \text{subject to} \quad h(x) - c = \phi_r(x)^\top A \phi_r(x) \quad \text{for all} \quad x \in \mathbb{R}^d$$

$$\phi_r(x) = (1, x_1, \ldots, x_d, x_1^2, x_1 x_2, x_1 x_3, \ldots, x_2^2, \ldots, x_1^r, \ldots, x_d^r) \in \mathbb{R}^n \qquad n = \binom{d+r}{r}$$

- Converges exactly when $h(x) - c^*$ is a sum of squares of polynomials
- Only few non-negative polynomials are a sum of squares
- Converges in the limit when $r \to \infty$. Slot, Laurent 2020 shows $n = O(\varepsilon^{-d/2})$

We use a larger space of models: algebraic property -> mild geometric condition.

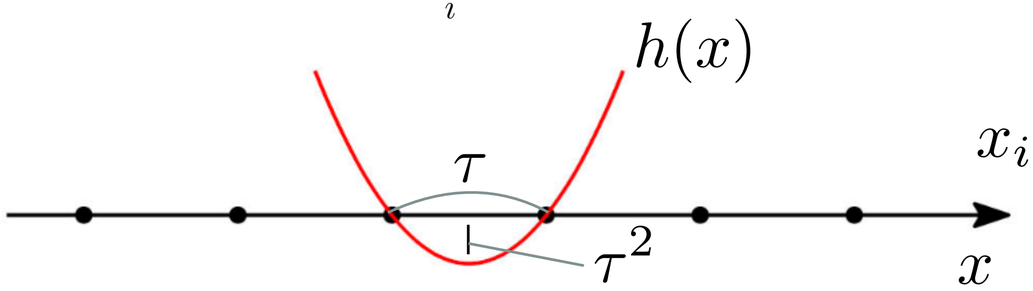
Bach, Rudi 2022: Polynomial hierarchy convergence $n = O((\log \frac{1}{\varepsilon})^d)$

Wait... why don't we just discretize?

$\hat{c} = \max_{c \in \mathbb{R}} c$ subject to $h(x_i) - c \geq 0, \quad \forall i = 1, \ldots, n$

Fancy way to write $\hat{c} = \min_i h(x_i)$

Example:



x_i grid of step τ

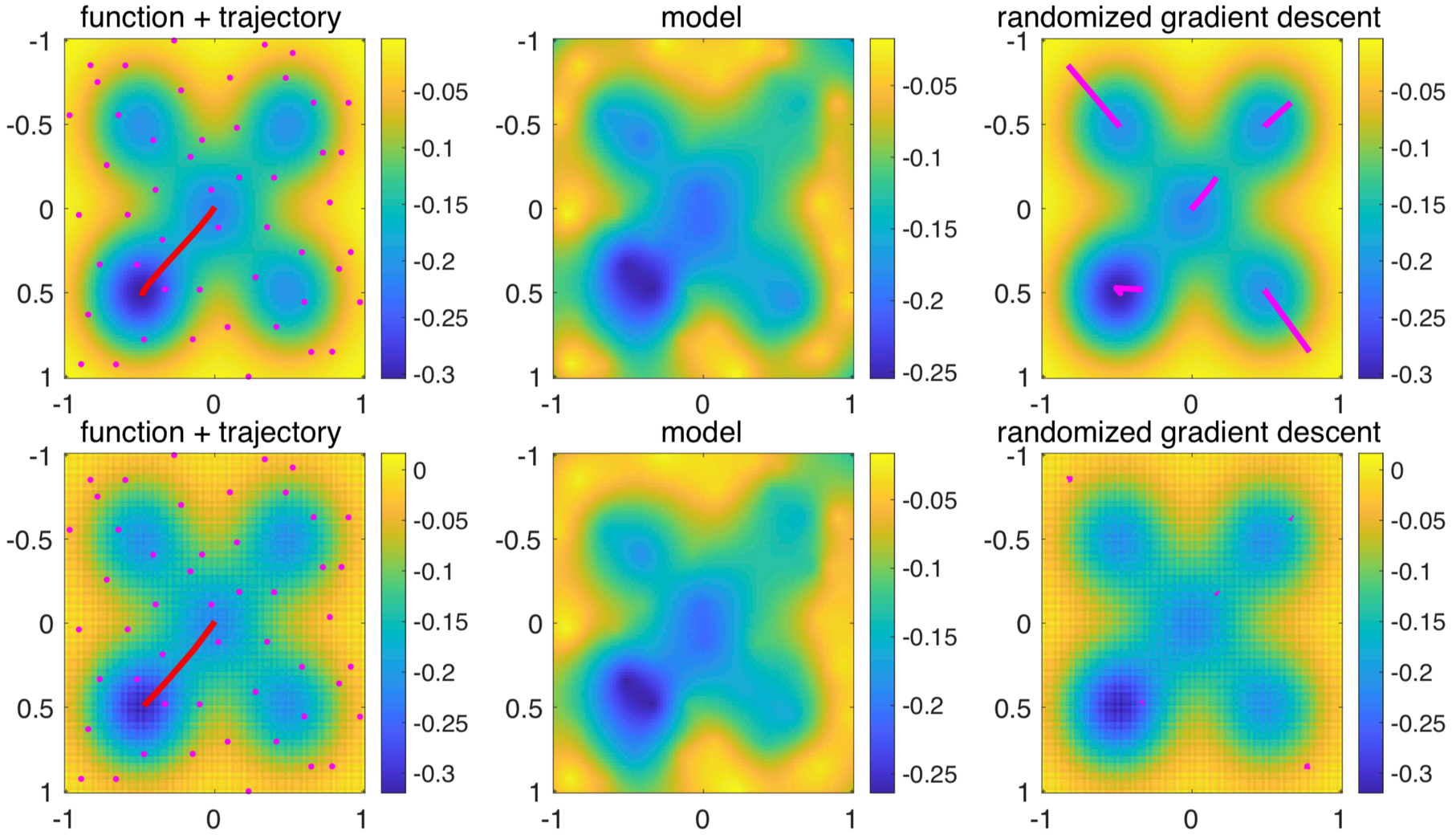
By construction $\hat{c} - c^* \geq O(\tau^2)$. Since $n = O(\tau^{-d})$, then

$n = O(\varepsilon^{-d/2})$ is necessary to achieve $|\hat{c} - c^*| \leq \varepsilon$

Discretizing alone is not enough!

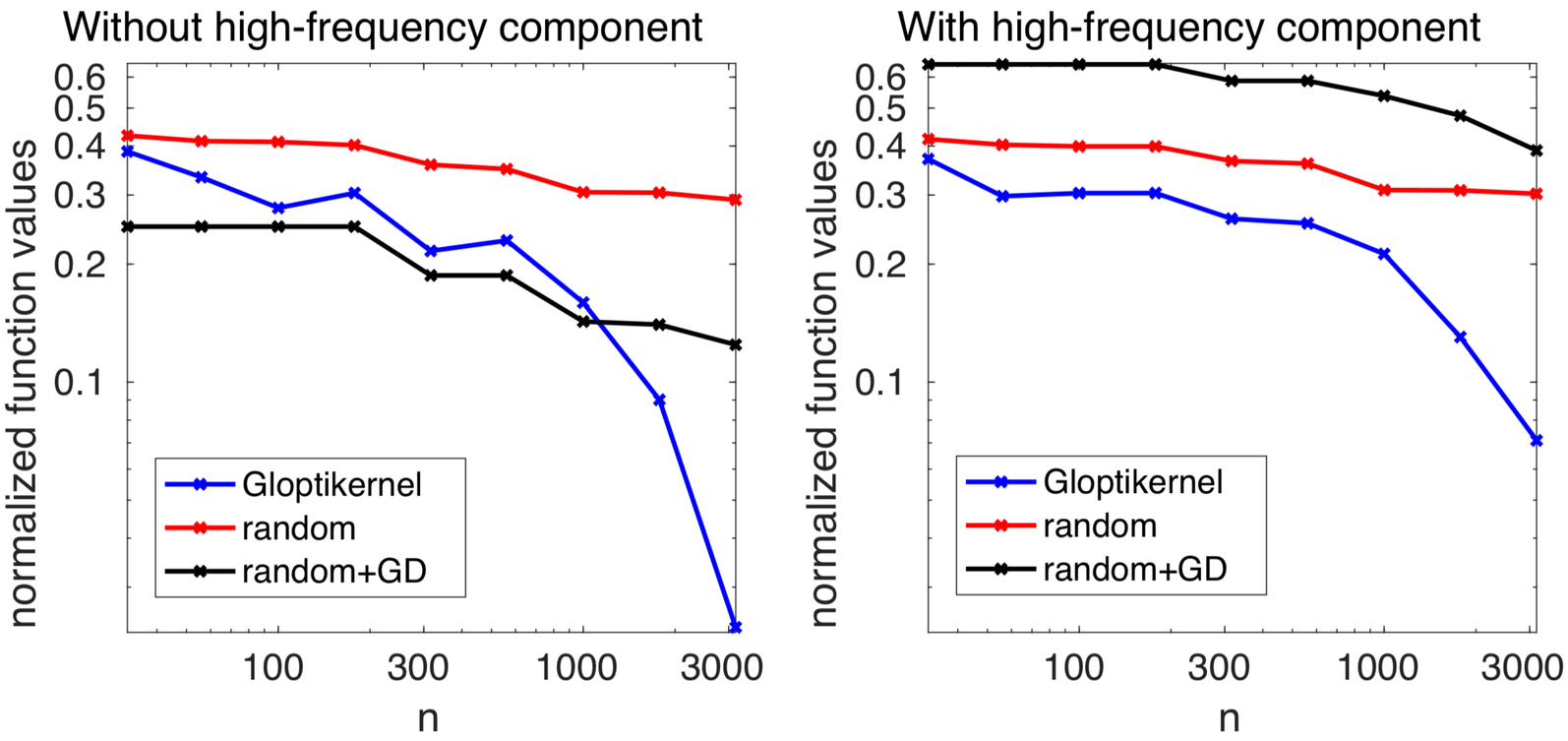
Example

d = 2


function + trajectory
model
randomized gradient descent
function + trajectory
model
randomized gradient descent

Example

d = 8


Without high-frequency component
normalized function values
0.6
0.5
0.4
0.3
0.2
0.1
100
300
1000
3000
n
Gloptikernel
random
random+GD
With high-frequency component
normalized function values
0.6
0.5
0.4
0.3
0.2
0.1
100
300
1000
3000
n
Gloptikernel
random
random+GD

Thank you!

Grand challenge…

	Proof of Concept	— Best on the field	— Established approach
• Algorithms	simplest according to theory	Explicit guarantees Few parameters with clear role Fast, scalable on big data Adaptive to the architecture	**LAPACK style library** **(High performance - standardised)**
• Applications	toy examples	Determine “the” benchmark Show that the method hammers it	**Real Problems** **(Nature, Science)**
• Modeling	Non-engineered model (Just from theory)	Specific sub-problems where we can do way better	Highly engineered, flexible model
• Theory	Enough to show that The method is promising	Refinement/simplification of the theory	General theory (All the elements in the right place) Open Questions