# Modelling functions with kernels, from logistic regression to global optimization


Ulysse Marteau-Ferey

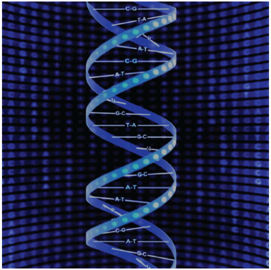
DI ENS – Inria Paris – PSL


Supervised by Francis Bach and Alessandro Rudi

September 7, 2022

# Introduction : learning a prediction function

Goal - Input $x \in \mathcal{X}$ $\xrightarrow{\text{predict}}$ output $y \in \mathcal{Y}$



$\longrightarrow$

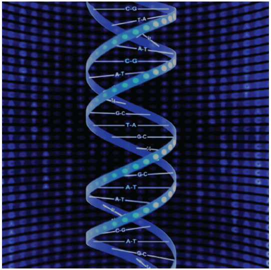
$\mathcal{Y} = \{\text{Healthy}, \text{Sick}\}$

$\mathcal{Y} = \{\text{Cancer A}, \text{Cancer B}\}$

$\mathcal{Y} = \{-1, 1\}$

# Introduction : learning a prediction function

Goal - Input $x \in \mathcal{X}$ $\overset{\text{predict}}{\longrightarrow}$ output $y \in \mathcal{Y}$



$\longrightarrow$

$$\mathcal{Y} = \{\text{Healthy}, \text{Sick}\}$$
$$\mathcal{Y} = \{\text{Cancer A}, \text{Cancer B}\}$$
$$\mathcal{Y} = \{-1, 1\}$$

Mathematically - Learning a **prediction function**

$$g : \mathcal{X} \to \mathcal{Y}$$

# Introduction : modelling

Mathematical formulation -

- Where to find $g$ : **model** $\mathcal{H}$ (set of test functions).
- How to find the best $g$ : minimize a risk

$$g_* = \underset{g \in \mathcal{H}}{\operatorname{argmin}} \mathcal{R}(g).$$

# Introduction : modelling

Mathematical formulation -

- Where to find $g$ : **model** $\mathcal{H}$ (set of test functions).
- How to find the best $g$ : minimize a risk

$$g_* = \underset{g \in \mathcal{H}}{\operatorname{argmin}} \mathcal{R}(g).$$

The model $\mathcal{H}$ is crucial -

- Large enough (good candidates).
- Small enough (not too many candidates).
- Impacts optimization (finding $g_*$).

# The linear model

Linear model -

- Features : $(\phi_i)_{1 \leq i \leq p}$, $\phi_i : \mathcal{X} \to \mathbb{R}$.
- Predictor :

$$g_\theta(x) = \sum_{j=1}^{p} \theta_j \, \phi_j(x) = \theta^\top \phi(x).$$

- Model :

$$\mathcal{H} = \{g_\theta \ : \ \theta \in \mathbb{R}^p\}$$

# The linear model

Linear model -

- Features : $(\phi_i)_{1 \le i \le p}$, $\phi_i : \mathcal{X} \to \mathbb{R}$.
- Predictor :

$$g_\theta(x) = \sum_{j=1}^{p} \theta_j \ \phi_j(x) = \theta^\top \phi(x).$$

- Model :

$$\mathcal{H} = \{g_\theta \ : \ \theta \in \mathbb{R}^p\}$$

Great workhorse in applied mathematics -

- Practitioners : feature design (interpretability).
- Theoreticians : "simplicity" of computations.
- Convex optimization algorithms.

# Motivations of the thesis and outline

Focus of the thesis : kernel methods -

- Generalization of linear models
- Non parametric : less rigid than linear models (bigger spaces)
- Very strong theoretical tools

# Motivations of the thesis and outline

Focus of the thesis : kernel methods -

- Generalization of linear models
- Non parametric : less rigid than linear models (bigger spaces)
- Very strong theoretical tools

Goal : extend the use of this tool -

1. Kernel methods
2. Logistic regression
3. Global (non-convex) optimization

# Part I - Kernel methods

# Reproducing Kernel Hilbert Spaces : two points of views

Hilbert space $\mathcal{H}, \langle \cdot, \cdot \rangle$ of functions on $\mathcal{X}$ [Aronszajn, 1950],[Scholkopf and Smola, 2001].

# Reproducing Kernel Hilbert Spaces : two points of views

Hilbert space $\mathcal{H}, \langle \cdot, \cdot \rangle$ of functions on $\mathcal{X}$ [Aronszajn, 1950],[Scholkopf and Smola, 2001].

Function evaluations are continuous -

- feature map : $x \in \mathcal{X} \mapsto \phi(x) \in \mathcal{H}$ s.t. $g(x) = \langle g, \phi(x) \rangle$ : **reproducing property** ;
- associated **positive definite kernel :** $k(x, y) = \langle \phi(x), \phi(y) \rangle$ ;
- **reproducing** : $\langle g, k(x, \cdot) \rangle = g(x)$ since $\phi(x) = k(x, \cdot)$.

# Reproducing Kernel Hilbert Spaces : two points of views

Hilbert space $\mathcal{H}, \langle \cdot, \cdot \rangle$ of functions on $\mathcal{X}$ [Aronszajn, 1950],[Scholkopf and Smola, 2001].

Function evaluations are continuous -

- feature map : $x \in \mathcal{X} \mapsto \phi(x) \in \mathcal{H}$ s.t. $g(x) = \langle g, \phi(x) \rangle$ : **reproducing property** ;
- associated **positive definite kernel :** $k(x, y) = \langle \phi(x), \phi(y) \rangle$ ;
- **reproducing** : $\langle g, k(x, \cdot) \rangle = g(x)$ since $\phi(x) = k(x, \cdot)$.

Positive definite kernel $k$ -

- basic functions : $g(\cdot) = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot) \rightarrow$ set $\mathcal{H}_0$ ;
- scalar product : $\langle k(x, \cdot), k(y, \cdot) \rangle = k(x, y)$ ;
- Hilbert space : $\mathcal{H} = \overline{\mathcal{H}_0}^{\langle \cdot, \cdot \rangle}$.

# Examples of RKHS

- Polynomial functions of degree $\leq r$.
- **Sobolev spaces** (regularity $s > d/2$) on $\mathcal{X} \subset \mathbb{R}^d$ (Lipschitz continuous).

$$f \in W_2^s(\mathcal{X}) \text{ if } \forall |\alpha| \leq s,\ \partial^\alpha f \in L^2(\mathcal{X})$$

- **Gaussian kernel** (bandwidth $\sigma$) :

$$k_\sigma(x, x') = \exp(-\|x - x'\|^2 / 2\sigma^2),$$

- Kernel engineering : design problem specific kernels [Scholkopf and Smola, 2001].

# The kernel trick

Classical optimization problem -

$$\widehat{g}_\lambda = \underset{g \in \mathcal{H}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} f_i(g(x_i)) + \frac{\lambda}{2} \|g\|^2$$

**Theorem (Representer theorem [Cucker and Smale, 2002])**

- *$\widehat{g}_\lambda$ of the form $\sum_{i=1}^n \alpha_i k(x_i, \cdot)$ where $\alpha \in \mathbb{R}^n$.*

$$\widehat{\alpha} = \underset{\alpha \in \mathbb{R}^n}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} f_i([\mathbf{K}\alpha]_i) + \frac{\lambda}{2} \alpha^\top \mathbf{K} \alpha,$$

$$\mathbf{K} = (k(x_i, x_j))_{1 \le i,j \le n} \text{ kernel matrix.}$$

Kernel trick -

- Looking in a $n$ dimensional space is enough.
- $\mathcal{H}$ only appears through the kernel.

# The kernel trick 2.0

Classical optimization problem -

$$\widehat{g}_\lambda = \operatorname*{argmin}_{g \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} f_i(g(x_i)) + \frac{\lambda}{2} \|g\|^2$$

**Theorem (Representer theorem [Cucker and Smale, 2002])**

- *$\widehat{g}_\lambda$ of the form $\sum_{i=1}^n \alpha_i k(x_i, \cdot)$ where $\alpha \in \mathbb{R}^n$.*

$$\widehat{\alpha} = \operatorname*{argmin}_{\alpha \in \mathbb{R}^n} \tfrac{1}{n} \sum_{i=1}^{n} f_i([\mathbf{K}\alpha]_i) + \tfrac{\lambda}{2} \alpha^\top \mathbf{K}\alpha,$$

$$\mathbf{K} = (k(x_i, x_j))_{1 \le i,j \le n} \text{ kernel matrix.}$$

Kernel trick 2.0 (informal) -

- Looking in a $m \ll n$ dimensional space is enough.
- $\mathcal{H}$ only appears through the kernel.

# Nice properties, classical drawbacks

Properties -

- Non parametric (infinite dimensional) : good approximation properties [Micchelli, Xu, and Zhang, 2006],[Sriperumbudur, Fukumizu, and Lanckriet, 2011].
- Kernel trick : finite dimensional problem + only use kernel.
- Tools for theoretical analysis : [Blanchard and Mücke, 2018],[Rudi, Carratino and Rosasco, 2017],[Scholkopf and Smola, 2001],[Caponnetto and de Vito, 2007].

Classical drawbacks -

- Scaling for large $n$ ($n > 10^6$).
- Hard to choose $k$ (non-isotropic data).

# Part II - Kernel logistic regression : extending results from least squares

Works presented in this section -

- **Statistics**
  Ulysse Marteau-Ferey, Dmitrii Ostrovskii, Francis Bach, and Alessandro Rudi. Beyond least- squares : Fast rates for regularized empirical risk minimization through self-concordance. COLT, 2019.
- **Optimization**
  Ulysse Marteau-Ferey, Francis Bach, and Alessandro Rudi. Newton methods for ill- conditioned generalized self-concordant losses. NeurIPS, 2019.

# Setting : supervised learning (1)

- Data : $(x_1, y_1), ..., (x_n, y_n) \in (\mathcal{X} \times \mathbb{R})^n$ i.i.d. from $\rho$ **unknown**.
- Predictors : $g \in \mathcal{H}$ RKHS with kernel $k$.
- Loss : $\ell(y, g(x)) \in \mathbb{R}_+$ :

Ideal goal - Expected risk minimization

$$g_* = \underset{g \in \mathcal{H}}{\text{argmin}}\ \mathcal{R}(g) := \mathbb{E}_{X,Y\sim\rho}[\ell(g(X), Y)]$$

- **Well-specified** assumption : $g_* \in \mathcal{H}$ exists.
- Access to $\rho$ through $(x_1, y_1), ..., (x_n, y_n)$.

# Setting : supervised learning (2)

Ideal goal - Expected risk minimization

$$g_* = \operatorname*{argmin}_{g \in \mathcal{H}} \mathcal{R}(g) := \mathbb{E}_{X,Y \sim \rho}[\ell(g(X), Y)] \tag{1}$$

Approximating $g_*$ in practice - Empirical risk minimization (ERM) :

- Replace $\rho \leftarrow \widehat{\rho} = \sum_{i=1}^{n} \delta_{(x_i, y_i)}$ :

$$\widehat{g}_\lambda = \operatorname*{argmin}_{g \in \mathcal{H}} \underbrace{\frac{1}{n} \sum_{i=1}^{n} \ell(g(x_i), y_i)}_{\text{empirical risk}} + \underbrace{\frac{\lambda}{2} \|g\|_{\mathcal{H}}^2}_{\text{regularization}} .$$

- Need regularization $\lambda$.

# Motivation : understanding and efficiently solving ERM

Previous work - quadratic case (or Kernel Ridge Regression) : closed form solutions.

$$\ell(y, y') = \frac{1}{2}\|y - y'\|^2$$

Goal - **Logistic regression** (no closed form solutions)

$$\ell(y, y') = \log(1 + \exp(-yy'))$$

1. Statistics : $\mathcal{R}(\widehat{g}_{\lambda,n}) - \mathcal{R}(g_*) = \Theta(n, \lambda)$.
2. Optimization : computing $\widehat{g}_\lambda$.

Main tools -

- Key property of logistic : Generalized Self Concordance ([Bach, 2010]).
- Newton method type analysis.

# Previous work : general statistical analysis

Bias-variance decomposition - [Sridharan et al.,2009] (assumption $\ell$ is $L$-Lipschitz).

$$\mathcal{R}(\widehat{g}_\lambda) - \mathcal{R}(g_*) \leq \underbrace{\lambda \|g_*\|^2}_{\text{bias } b_\lambda} + \underbrace{\frac{L^2}{\lambda n}}_{\text{variance } d_\lambda}$$

- $b_\lambda$ : regularity of $g_*$
- $d_\lambda$ : effective dimension of the problem

Rates of convergence -

$$\mathcal{R}(\widehat{g}_\lambda) - \mathcal{R}(g_*) \leq \frac{L\|g_*\|}{\sqrt{n}}$$

**In practice : faster convergence. Why ?**

# Refined bias-variance decompositions

Bias-variance decomposition (non-asymptotic) -
Least squares : [Caponnetto and de Vito,2007],[Blanchard and Mücke, 2018]
Logistic (GSC functions) : [M-F., Ostrovskii, Bach and Rudi, 2019]

$$\mathcal{R}(\widehat{g}_\lambda) - \mathcal{R}(g_*) \leq b_\lambda + \frac{d_\lambda}{n},$$

# Refined bias-variance decompositions

Bias-variance decomposition (non-asymptotic) -
Least squares : [Caponnetto and de Vito,2007],[Blanchard and Mücke, 2018]
Logistic (GSC functions) : [M-F., Ostrovskii, Bach and Rudi, 2019]

$$\mathcal{R}(\widehat{g}_\lambda) - \mathcal{R}(g_*) \leq b_\lambda + \frac{d_\lambda}{n},$$

| | bias $b_\lambda$ | effective dimension $d_\lambda$ |
|---|---|---|
| $L$-lipschitz | $\lambda\|g_*\|^2$ | $L^2/\lambda$ |
| least squares | $\lambda^2\|(\Sigma+\lambda I)^{-1/2}g_*\|^2$ | $\mathrm{Tr}((\Sigma+\lambda I)^{-1}\Sigma)$ |
| logistic | $\lambda^2\|(H+\lambda I)^{-1/2}g_*\|^2$ | $\mathrm{Tr}((H+\lambda I)^{-1}G)$ |

- Least squares : covariance operator $\Sigma = \mathbb{E}[k_X \otimes k_X] \in \mathcal{S}_+(\mathcal{H})$
- GSC functions : Hessian and Fisher information operators at $g_*$ : $H, G$.

**Finer analysis : better understanding**

# Least squares : minimax optimal rates

Assumptions -

- $d_\lambda \asymp \lambda^{-1/b}$ for $b \geq 1$ ($b \uparrow$ if size of $\mathcal{H}$ decreases).
- $b_\lambda \asymp \lambda^{2r}$ for $r \in [1/2, 1]$ ($r \uparrow$ regularity of $g_*$).

Theorem ([Caponnetto and de Vito, 2007])

*Minimax upper and lower bounds :*

$$\mathcal{R}(\widehat{g}_\lambda) - \mathcal{R}(g_*) \asymp n^{-\frac{2br}{2br+1}}, \qquad \lambda \asymp n^{-\frac{b}{2br+1}}$$

| rate | $b = 1$ | $b \to +\infty$ |
|---|---|---|
| $r = 1/2$ | $n^{-1/2}$ | $n^{-1}$ |
| $r = 1$ | $n^{-2/3}$ | $n^{-1}$ |

**Much more precise result, and reflects behavior in practice**

# Logistic regression and GSC functions

Assumptions -

- $d_\lambda \lesssim \lambda^{-1/b}$ for $b \geq 1$ ($b \uparrow$ if size of $\mathcal{H}$ decreases).
- $b_\lambda \lesssim \lambda^{2r}$ for $r \in [1/2, 1]$ ($r \uparrow$ regularity of $g_*$).

Theorem ([M-F., Ostrovskii, Bach and Rudi (2019)])

*Non asymptotic upper bounds :*

$$\mathcal{R}(\widehat{g}_\lambda) - \mathcal{R}(g_*) \lesssim n^{-\frac{2br}{2br+1}}, \qquad \lambda \asymp n^{-\frac{b}{2br+1}}$$

| rate | $b = 1$ | $b \to +\infty$ |
|---|---|---|
| $r = 1/2$ | $n^{-1/2}$ | $n^{-1}$ |
| $r = 1$ | $n^{-2/3}$ | $n^{-1}$ |

**Much more precise result, and reflects behavior in practice**

# Main tool for logistic regression

Assumption/Tool - : Generalized Self-Concordance [Bach, 2010] (GSC, satisfied by logistic loss).

$$|\ell^{(3)}(\cdot, y_2)| \leq \ell^{(2)}(\cdot, y_2).$$

Useful consequences -

- Allows to **localize the minimum** when the Newton decrement is small enough :

$$\|(H + \lambda I)^{-1} \nabla \mathcal{R}(g)\| \leq r_\lambda \implies \|g - g_*\| \leq c.$$

- **Quadratic behavior** :

$$\|g - g_*\| \leq c \implies \mathcal{R}(g) - \mathcal{R}(g_*) \leq b_\lambda + C(\lambda)\|g - g_*\|_H^2.$$

# Finite dimensional ERM

$$\widehat{g}_\lambda = \underset{g \in \mathcal{H}}{\operatorname{argmin}} \ \frac{1}{n} \sum_{i=1}^{n} \ell(g(x_i), y_i) + \frac{\lambda}{2} \|g\|^2$$

Kernel trick : finite dimensional problem-

$$\widehat{g}_\lambda(\cdot) = \sum_{i=1}^{n} \widehat{\alpha}_i k(x_i, \cdot),$$

$$\widehat{\alpha} = \underset{\alpha \in \mathbb{R}^n}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \ell([K\alpha]_i, y_i) + \frac{\lambda}{2} \alpha^\top K \alpha,$$

$$K = (k(x_i, x_j))_{1 \leq i,j \leq n}.$$

Linear system in the quadratic case -

$$(K + \lambda I)\widehat{\alpha} = Ky.$$

# Fast kernel ridge regression

Key observation - Only statistical precision is needed.

Fast, scalable algorithm (FALKON, [Rudi et al., 2017])

Main techniques -

- Nyström projections ([Rudi et al. 2015]) : reduce dimension from $n$ to $m = d_\lambda \ll n$
- Pre-conditioning + iterative method.

**Theorem** (Rudi et al.,2017)

*There exists an algorithm which achieves statistical optimality in* $O(nd_\lambda + d_\lambda^3)$ ***in time and*** $O(n)$ ***in space***

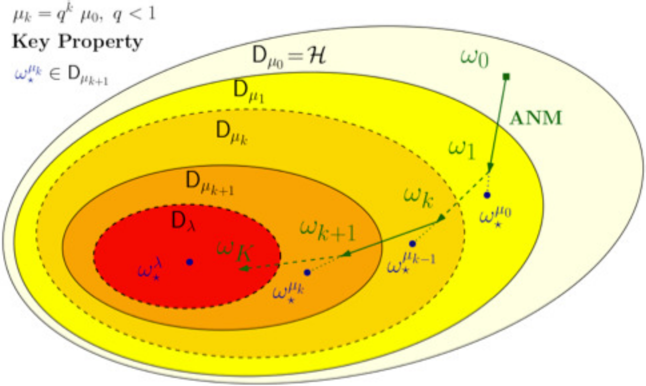
Scalable to large datasets - $n = 10^9$ points, $\lambda$ small.

# Extension to logistic regression

Key observations -

- Previous techniques : fast approximate Newton steps.
- Empirically, Newton converges for logistic.

Globally convergent Newton method - Regularization $\mu_k \downarrow \lambda$



# Extension to logistic regression

Key observations -

- Previous techniques : fast approximate Newton steps.
- Empirically, Newton converges for logistic.

Globally convergent Newton method -

- Decrease regularization $\mu \downarrow \lambda$ linearly.

**Theorem (M-F., Bach, Rudi, 2019)**

*There exists an algorithm which reaches the statistical upper bound in* $O(\log(1/\lambda)(nd_\lambda + d_\lambda^3))$ ***in time and*** $O(n)$ ***in space***

# Conclusion on logistic regression

The strength of second order methods -

- First order methods depend on **condition number** $\kappa = \frac{L}{\lambda}$.
- Small $\lambda$ sometimes necessary (Higgs data set : $\lambda = 10^{-12}$).
- Second order methods : no (or logarithmic) dependence in $\kappa$.
- Good non-asymptotic analysis tool.

# Conclusion on logistic regression

The strength of second order methods -

- First order methods depend on **condition number** $\kappa = \frac{L}{\lambda}$.
- Small $\lambda$ sometimes necessary (Higgs data set : $\lambda = 10^{-12}$).
- Second order methods : no (or logarithmic) dependence in $\kappa$.
- Good non-asymptotic analysis tool.

Related works -

- [Beugnot, Mairal and Rudi, 2021] Theory : statistical rates for $r \geq 1$ (very smooth, not ERM).
- [Meanti et al., 2020] Experimental : library for this work and FALKON (up to $n = 10^9$).

# Part III - Global (non-convex) optimization

Main work presented in this section -

- Alessandro Rudi, Ulysse Marteau-Ferey, and Francis Bach. Finding global minima via kernel approximations. Arxiv, 2020.

Other works used in this section -

- PSD models.
  Ulysse Marteau-Ferey, Francis Bach, and Alessandro Rudi. Non-parametric models for non-negative functions. NeurIPS, 2020.
- Extension to manifolds.
  Ulysse Marteau-Ferey, Francis Bach, and Alessandro Rudi. Second order conditions to decompose smooth functions as sums of squares. Arxiv, 2022.

# Global non-convex optimization : setting

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 regularity (here $f \in C^m(\Omega)$)
- access to function calls (no derivatives)
- **no convexity assumption**

# Global non-convex optimization : setting

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 regularity (here $f \in C^m(\Omega)$)
- access to function calls (no derivatives)
- **no convexity assumption**

Goal - Given $\varepsilon > 0$, find $\hat{x} \in \Omega$ such that $f(\hat{x}) - \min_{x \in \Omega} f(x) \leqslant \varepsilon$

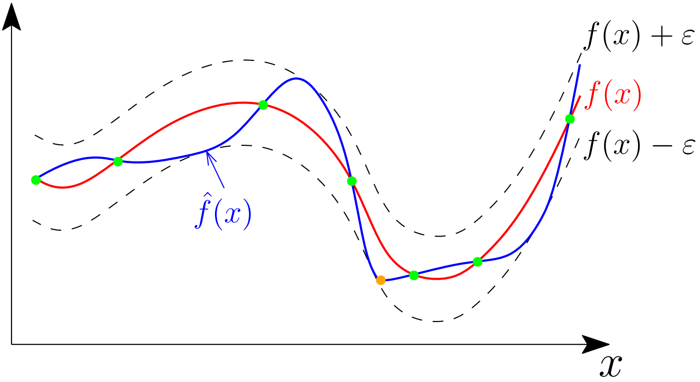
- Lowest number of function calls $n$ ;
- Worst-case guarantees over all functions $f$ in $C^m(\Omega)$

$$\sup_{f \in C^m(\Omega),\ \|f\| \leq B} \left\{ f(\hat{x}) - \min_{x \in \Omega} f(x) \right\} \leqslant \varepsilon$$

# Optimal algorithms

Goal - Given $\varepsilon > 0$, find $\widehat{x} \in \Omega$ such that $f(\widehat{x}) - \min_{x \in \Omega} f(x) \leqslant \varepsilon$.

Equivalent to **uniform function approximation** [Novak, 2006].
Simplest algorithm : approximate $f$ by $\widehat{f}$ and minimize $\widehat{f}$.



# Optimal rates

Optimal worst-case performance over $C^m$ - [Novak,2006]

- $n =$ number of function evaluations needed ;
- $m = 1$, $n \propto \varepsilon^{-d}$ : **curse of dimensionality** ;

# Optimal rates

Optimal worst-case performance over $C^m$ - [Novak,2006]

- $n =$ number of function evaluations needed ;
- $m = 1$, $n \propto \varepsilon^{-d}$ : **curse of dimensionality** ;
- $m$ bounded derivatives : $n \propto \varepsilon^{-d/m}$.
- NB : constants may depend (exponentially) in $d$

# Optimal rates

Optimal worst-case performance over $C^m$ - [Novak,2006]

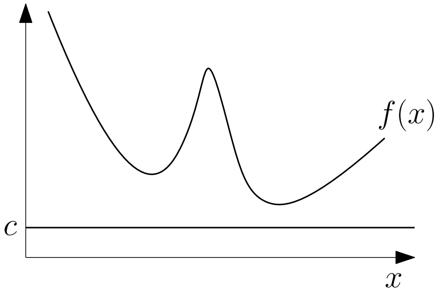
- $n =$ number of function evaluations needed ;
- $m = 1$, $n \propto \varepsilon^{-d}$ : **curse of dimensionality** ;
- $m$ bounded derivatives : $n \propto \varepsilon^{-d/m}$.
- NB : constants may depend (exponentially) in $d$

Algorithms -

- Current algorithms have **exponential** running time complexity in $n$ : "approximate then optimize".
- Algorithms with **polynomial-time** complexity in $n$ : “approximate **and** optimize” ?

# Reformulation : all optimization problems are convex

$$\min_{x\in\Omega} f(x) = \sup_{c\in\mathbb{R}} \quad c$$
$$\text{subject to} \quad f(x) - c = g(x),$$
$$g(x) \geq 0, \ x \in \Omega$$

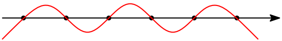


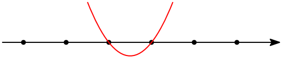

Need to **represent non-negative functions** (such as $g = f - c$)

# PSD strengthening

Motivations -

- Constraint $g \geq 0$ hard.
- Discretizing $g \geq 0$ as $g(x_i) \geq 0$ does not leverage regularity $\implies$ approximate the whole of $g$ ?





- Approximate the whole of $g$ ?

# PSD strengthening

Motivations -

- Constraint $g \geq 0$ hard.
- Discretizing $g \geq 0$ as $g(x_i) \geq 0$ does not leverage regularity $\implies$ approximate the whole of $g$ ?

- Approximate the whole of $g$ ?

One possible solution : PSD strengthening -

- Feature map $\phi : \mathcal{X} \to \mathcal{H}$
- Parametrized by **positive semi-definite operators**

$$g_A(x) = \langle \phi(x), A\phi(x) \rangle, \; A \in S_+(\mathcal{H}).$$

- Enforces non-negativity structurally ($A \succeq 0 \implies g_A \geq 0$) while being **linear** in $A$.

# Kernel PSD models and sum of squares

$$g_A(x) = \langle \phi(x), A\phi(x) \rangle, \; A \in S_+(\mathcal{H}).$$

Kernel PSD models/sum of squares - [M-F., Bach, Rudi, 2020]

- Use $\phi(x) = k_x = k(x, \cdot)$ where $k$ is a positive definite kernel.
- Spectral theorem : $g_A$ are **sum of squares** of functions in $\mathcal{H}$.
- **Here** $k_s$ kernel associated to $W_2^s(\Omega)$.

# Kernel PSD models and sum of squares

$$g_A(x) = \langle \phi(x), A\phi(x) \rangle, \; A \in S_+(\mathcal{H}).$$

Kernel PSD models/sum of squares - [M-F., Bach, Rudi, 2020]

- Use $\phi(x) = k_x = k(x, \cdot)$ where $k$ is a positive definite kernel.
- Spectral theorem : $g_A$ are **sum of squares** of functions in $\mathcal{H}$.
- **Here** $k_s$ kernel associated to $W_2^s(\Omega)$.

New problem -

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

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

Step 1 - Showing the strengthening is reached for some $k = k_s$, $s > d/2$

$$\sup_{c \in \mathbb{R},\ A \succeq 0} c \quad \text{st} \quad \forall x \in \Omega,\ f(x) - c = \langle k_x, A k_x \rangle \tag{2}$$

$$\sup_{c \in \mathbb{R}} c \quad \text{st} \quad \forall x \in \Omega,\ f(x) - c \geq 0 \tag{3}$$

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

Step 1 - Showing the strengthening is reached for some $k = k_s$, $s > d/2$

$$\sup_{c \in \mathbb{R},\ A \succeq 0} \quad c \quad \text{st} \ \ \forall x \in \Omega,\ \ f(x) - c = \langle k_x, A k_x \rangle \tag{2}$$

$$\sup_{c \in \mathbb{R}} \ c \quad \text{st} \ \ \forall x \in \Omega,\ \ f(x) - c \geq 0 \tag{3}$$

**Condition** :

$$\exists A_* \in S_+(\mathcal{H}),\ \ f(x) = f_* + \langle k_x, A_* k_x \rangle$$

$f - f_*$ can be written as a sum of functions in $W_2^s(\Omega)$

$$f - f_* = \sum_{i=1}^{N} f_i^2, \qquad f_i \in W_2^s(\Omega)$$

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

Step 1 - Showing the strengthening is tight for some $k = k_s$, $s > d/2$

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

Step 2 - Discretize using $n$ evaluations points $(x_i)_{1 \leq i \leq n}$ :

$$\widehat{c}, \widehat{A} = \underset{c \in \mathbb{R},\ A \in S_+(\mathcal{H})}{\operatorname{argmax}} \quad c - \lambda \operatorname{Tr}(A) \tag{4}$$
$$\text{subject to} f(x_i) - c = \langle k_{x_i}, A k_{x_i} \rangle,\ 1 \leq i \leq n$$

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

Step 1 - Showing the strengthening is tight for some $k = k_s$, $s > d/2$

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

Step 2 - Discretize using $n$ evaluations points $(x_i)_{1 \leq i \leq n}$ :

$$\begin{aligned} \widehat{c}, \widehat{A} = & \underset{c \in \mathbb{R},\ A \in S_+(\mathcal{H})}{\operatorname{argmax}} \quad c - \lambda \operatorname{Tr}(A) \\ & \text{subject to} f(x_i) - c = \langle k_{x_i}, A k_{x_i} \rangle,\ 1 \leq i \leq n \end{aligned} \tag{4}$$

- $n$ large enough to guarantee that $\|\widehat{c} - f_*\| \leq \epsilon$
- regularization $\lambda \operatorname{Tr}(A)$ necessary to avoid overfitting

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

Step 1 - Showing the strengthening is tight

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

Step 2 - Discretize using $n$ evaluations points $(x_i)_{1 \le i \le n}$ :

$$\begin{aligned} \widehat{c}, \widehat{A} = & \underset{c \in \mathbb{R},\ A \in S_+(\mathcal{H})}{\operatorname{argmax}} \quad c - \lambda \operatorname{Tr}(A) \\ & \text{subject to } f(x_i) - c = \langle k_{x_i}, A k_{x_i} \rangle,\ 1 \le i \le n \end{aligned} \tag{4}$$

Step 3 - Show (4) can be written as a $n \times n$ semidefinite program.

- Consequence of the representer theorem
- Solve with interior point methods $O(n^{3.5})$ [Nesterov and Nemirovskii, 1994],[Tuncel, 2004]
- Dimension reduction (Nyström, [Rudi et al.,2015])

# Step 1 : tight strengthening

**Theorem ([Rudi, M-F., Bach, 2020])**

*Assume $\Omega$ is bounded, $f \in C^m(\Omega)$ has* ***isolated strict-second order minima****, and that $\{f - f_* \leq \delta\} \subset \mathring{\Omega}$ for some $\delta > 0$.*
*For any $s \in ]d/2, m-2]$, there exists $h_1, \dots, h_N \in W_2^s(\Omega)$ such that*

$$\forall x \in \Omega,\ f(x) = f_* + \sum_{i=1}^{N} h_i^2(x)$$
$$= f_* + \langle k_x, A_* k_x \rangle_{\mathcal{H}} \text{ where } A_* = \sum h_i \otimes h_i$$

- Analog of Positivstellensatz for the polynomial case ([Putinar, 1993],[Lasserre, 2010]).
- Manifolds and continuous sets of minima [M-F., Bach, Rudi, 2022], motivated by [Vacher et al.].

# Step 2 : discretizing using random samples

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

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

**Theorem ([Rudi,M-F.,Bach, 2020])**

*Up to logarithmic terms : $x_1, \dots, x_n$ sampled uniformly from $\Omega$. Up to log terms, if $n = O(\varepsilon^{-d/(m-d/2-3)})$, $\lambda = \varepsilon$, then it holds with probability at least $1 - \delta$ :*

$$|\widehat{c} - f_*| \leq \varepsilon \ \operatorname{Tr}(A_*) \ \log \frac{1}{\delta}$$

- Near optimal ($\varepsilon^{-d/(m-d/2)}$ for Sobolev).
- In practice, $n$ is a computational budget.

# Step 3 : Pseudocode for the 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 $\Omega$
2. **Feature computation**
   - Set $f_j = f(x_j)$, $\forall j \in \{1, \dots, n\}$
   - Compute $K_{ij} = k_s(x_i, x_j)$
   - Set $\Phi_j \in \mathbb{R}^n$ computed using a Cholesky decomposition of $K$ $\forall j \in \{1, \dots, n\}$.
3. **Solve**

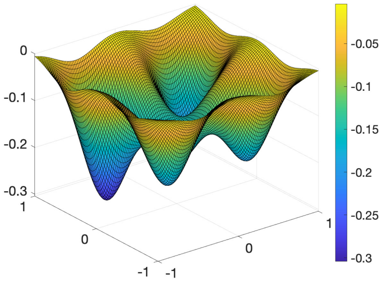
$$\max_{c \in \mathbb{R}, B \succcurlyeq 0} c - \lambda \operatorname{Tr}(B) \ \text{ s. t. } \ \forall j \in \{1, \dots, n\}, f_j - c = \Phi_j^\top B \Phi_j$$

**Output :** $c$ proxy for $f_*$.

Extension to compute $\widehat{x}$ possible

# First experiments

Example of function -




Experiments on benchmarks -

| | d | error |
|---|---|---|
| Trid | 6 | 0.00E+00 |
| Watson | 6 | 1.09E-03 |
| Hartmann6 | 6 | 0.00E+00 |
| LennardJones | 6 | 0.00E+00 |
| Thurber | 7 | 9.70E+03 |
| Xor | 9 | 6.99E-03 |
| Paviani | 10 | 1.03E-04 |
| Cola | 17 | 3.35E-01 |

# Optimization using sum of squares polymials

Polynomial sum of squares - $f$ is a polynomial [Lasserre,2001]

$$\rho_r = \sup_{c \in \mathbb{R}} \; c \quad \text{st} \quad f - c \in \Sigma_r[\boldsymbol{x}]$$

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

- $\phi(x) = (x^\alpha)_{|\alpha| \leq r}$.
- Optimization on a semi-algebraic domain $\mathbb{K}$ :

$$\mathbb{K} := \{g_i(x) \geq 0 \;:\; g_i \in \mathbb{R}[\boldsymbol{x}]\}$$

- More general framework : moment-SOS hierarchies (of lower bounds).

# Parallel with moment-sos hierarchy

Moment-sos hierarchy -

- Polynomials on semi-algebraic sets
- Guarantees based on **algebraic properties**
- A priori guarantees on the degree needed for a given precision ($r = 1/\sqrt{\varepsilon}$)
- SDP problem of dimension $d^r$
- A posteriori lower bounds
- exact extraction

Kernel Sum of Squares -

- Any function $f$ (but no constraints)
- Guarantees based on **regularity**
- A priori guarantees on the number of samples $n$ needed for a given precision ($n = \varepsilon^{-d/(m-d/2)}$)
- SDP of dimension $n$

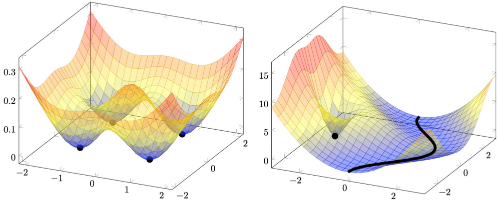
# Summary of global non-convex optimization

Takeaways -

- Algorithm for global optimization with $n$ evaluation points **polynomial** in $n$ (SDP).
- Guarantees for smooth functions : error $\varepsilon$ roughly $n = O(\varepsilon^{-d/(m-d/2)})$ points.

Related works -

- A posteriori guarantees with Fourier transform [Woodworth, Bach, Rudi, 2022].
- Set of minima is a sub-manifold of a manifold [M-F., Bach, Rudi, 2022].



# Open questions and future work directions

Logistic regression -

- Analysis of first order methods with GSC ?
- Upper rates for the misspecified setting.

Global optimization and sum of squares -

- Can constraints be added in global optimization ? What are their impact ?
- Finding a posteriori guarantees in certain interesting cases.
- Creation of a library.
- Models for shape constraints (outputs in the simplex or in a box for example).

Thank you for your attention !