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 $x \in \mathcal{X}$ predict output $y \in \mathcal{Y}$

- $\mathcal{Y} = \{\text{Healthy, Sick}\}$
- $\mathcal{Y} = \{\text{Cancer A, Cancer B}\}$
- $\mathcal{Y} = \{-1, 1\}$
Introduction: learning a prediction function

Goal - Input $x \in \mathcal{X}$ \hspace{1cm} predict \hspace{1cm} output $y \in \mathcal{Y}$

Mathematically - Learning a prediction function

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

\[ \mathcal{Y} = \{ \text{Healthy, Sick} \} \]
\[ \mathcal{Y} = \{ \text{Cancer A, Cancer B} \} \]
\[ \mathcal{Y} = \{ -1, 1 \} \]
Introduction: modelling

Mathematical formulation -

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

$$g_* = \arg\min_{g \in \mathcal{H}} R(g).$$
Introduction: modelling

Mathematical formulation -

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

$$g_* = \arg\min_{g \in \mathcal{H}} R(g).$$

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

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

- 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 : \((\phi_i)_{1 \leq i \leq p}, \phi_i : \mathcal{X} \rightarrow \mathbb{R}\).
- Predictor :
  \[g_\theta(x) = \sum_{j=1}^{p} \theta_j \phi_j(x) = \theta^\top \phi(x)\].
- Model :
  \[\mathcal{H} = \{g_\theta \mid \theta \in \mathbb{R}^p\}\]

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

1. Kernel methods
2. Logistic regression
3. Global optimization
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 $\phi(x) = k(x, \cdot)$. 
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: $g(\cdot) = \sum_{i=1}^{n} \alpha_{i} k(x_{i}, \cdot) \to \text{set } \mathcal{H}_{0}$;
- scalar product: $\langle k(x, \cdot), k(y, \cdot) \rangle = k(x, y)$;
- Hilbert space: $\mathcal{H} = \mathcal{H}_{0}^{\langle \cdot, \cdot \rangle}$. 
Examples of RKHS

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

\[ f \in W^s_2(\mathcal{X}) \text{ if } \forall |\alpha| \leq s, \; \partial^\alpha f \in L^2(\mathcal{X}) \]

- **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 -

\[ \hat{g}_\lambda = \arg\min_{g \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} f_i(g(x_i)) + \frac{\lambda}{2} \|g\|^2 \]

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

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

\[ \hat{\alpha} = \arg\min_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} f_i([K\alpha]_i) + \frac{\lambda}{2} \alpha^\top K \alpha, \]

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

\[ \hat{g}_\lambda = \arg\min_{g \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} f_i(g(x_i)) + \frac{\lambda}{2} \| g \|^2 \]

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

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

\[ \hat{\alpha} = \arg\min_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} f_i([K\alpha]_i) + \frac{\lambda}{2} \alpha^\top K \alpha, \]

\[ K = (k(x_i, x_j))_{1 \leq i, j \leq n} \text{ kernel matrix}. \]

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.

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**

- **Optimization**
Setting: supervised learning (1)

- Data: \((x_1, y_1), \ldots, (x_n, y_n) \in (\mathcal{X} \times \mathbb{R})^n\) 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

\[
g_\ast = \arg\min_{g \in \mathcal{H}} \mathcal{R}(g) := \mathbb{E}_{X, Y \sim \rho}[\ell(g(X), Y)]
\]

- **Well-specified** assumption: \(g_\ast \in \mathcal{H}\) exists.
- Access to \(\rho\) through \((x_1, y_1), \ldots, (x_n, y_n)\).
Ideal goal - Expected risk minimization

\[ g_\star = \arg\min_{g \in \mathcal{H}} \mathcal{R}(g) := \mathbb{E}_{X,Y \sim \rho}[\ell(g(X), Y)] \tag{1} \]

Approximating \( g_\star \) in practice - Empirical risk minimization (ERM) :

\begin{itemize}
  \item Replace \( \rho \leftarrow \hat{\rho} = \sum_{i=1}^{n} \delta(x_i, y_i) \) :
  \[ \hat{g}_\lambda = \arg\min_{g \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(g(x_i), y_i) + \frac{\lambda}{2} \|g\|_H^2 \]
  \item Need regularization \( \lambda \).
\end{itemize}
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: \( R(\hat{g}_\lambda, n) - R(g_\star) = \Theta(n, \lambda) \).
2. Optimization: computing \( \hat{g}_\lambda \).

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).

$$\mathcal{R}(\hat{g}_\lambda) - \mathcal{R}(g_*) \leq \lambda \|g_*\|^2 + \frac{L^2}{\lambda n}$$

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

Rates of convergence -

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

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]

\[ \mathcal{R}(\hat{g}_\lambda) - \mathcal{R}(g_*) \leq b_\lambda + \frac{d_\lambda}{n}, \]
# 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]

![Math equation]

<table>
<thead>
<tr>
<th></th>
<th>bias $b_\lambda$</th>
<th>effective dimension $d_\lambda$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L$-lipschitz</td>
<td>$\lambda |g_\lambda|^2$</td>
<td>$L^2/\lambda$</td>
</tr>
<tr>
<td>least squares</td>
<td>$\lambda^2 |(\Sigma + \lambda I)^{-1/2}g_\lambda|^2$</td>
<td>$\text{Tr}((\Sigma + \lambda I)^{-1}\Sigma)$</td>
</tr>
<tr>
<td>logistic</td>
<td>$\lambda^2 |(H + \lambda I)^{-1/2}g_\lambda|^2$</td>
<td>$\text{Tr}((H + \lambda I)^{-1}G)$</td>
</tr>
</tbody>
</table>

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

**Finer analysis : better understanding**
Least squares: minimax optimal rates

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

**Theorem ([Caponnetto and de Vito, 2007])**

**Minimax upper and lower bounds:**

\[
\mathcal{R}(\hat{g}_λ) - \mathcal{R}(g_*) \asymp n^{-\frac{2br}{2br+1}}, \quad \lambda \asymp n^{-\frac{b}{2br+1}}
\]

<table>
<thead>
<tr>
<th>rate</th>
<th>( b = 1 )</th>
<th>( b \to +\infty )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( r = 1/2 )</td>
<td>( n^{-1/2} )</td>
<td>( n^{-1} )</td>
</tr>
<tr>
<td>( r = 1 )</td>
<td>( n^{-2/3} )</td>
<td>( n^{-1} )</td>
</tr>
</tbody>
</table>

Much more precise result, and reflects behavior in practice
Logistic regression and GSC functions

Assumptions -

1. $d_\lambda \lesssim \lambda^{-1/b}$ for $b \geq 1$  
   ($b \uparrow$ if size of $\mathcal{H}$ decreases).

2. $b_\lambda \lesssim \lambda^{2r}$ for $r \in [1/2, 1]$  
   ($r \uparrow$ regularity of $g^*$).

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

*Non asymptotic upper bounds:*

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

<table>
<thead>
<tr>
<th>r</th>
<th>rate</th>
<th>$b = 1$</th>
<th>$b \to +\infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1/2$</td>
<td>$n^{-1/2}$</td>
<td>$n^{-1}$</td>
<td></td>
</tr>
<tr>
<td>$1$</td>
<td>$n^{-2/3}$</td>
<td>$n^{-1}$</td>
<td></td>
</tr>
</tbody>
</table>

Much more precise result, and reflects behavior in practice
Main tool for logistic regression


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

Useful consequences -

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

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

- **Quadratic behavior**:

\[ \|g - g_*\| \leq c \implies R(g) - R(g_*) \leq b_\lambda + C(\lambda)\|g - g_*\|^2_H. \]
**Finite dimensional ERM**

\[
\hat{g}_\lambda = \arg\min_{g \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(g(x_i), y_i) + \lambda \frac{1}{2} \|g\|^2
\]

Kernel trick: finite dimensional problem-

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

\[
\hat{\alpha} = \arg\min_{\alpha \in \mathbb{R}^n} \frac{1}{n} \sum_{i=1}^{n} \ell([K\alpha]_i, y_i) + \frac{1}{2} \alpha^\top K \alpha,
\]

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

Linear system in the quadratic case -

\[
(K + \lambda I) \hat{\alpha} = Ky.
\]
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 $O(\log(1/\lambda)(nd_\lambda + d^3_\lambda))$ 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 -


Other works used in this section -

- PSD models.

- Extension to manifolds.
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
\[
\left\{ f(\hat{x}) - \min_{x \in \Omega} f(x) \right\} \leq \varepsilon
\]
- Lowest number of function calls \( n \);
- Worst-case guarantees over all functions \( f \) in \( C^m(\Omega) \)

\[
\sup_{f \in C^m(\Omega), \|f\| \leq B} \left\{ f(\hat{x}) - \min_{x \in \Omega} f(x) \right\} \leq \varepsilon
\]
Optimal algorithms

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

Equivalent to **uniform function approximation** [Novak, 2006].

Simplest algorithm: approximate $f$ by $\hat{f}$ and minimize $\hat{f}$.
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

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

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

One possible solution: PSD strengthening -
- Feature map $\phi: \mathcal{X} \rightarrow \mathcal{H}$
- Parametrized by positive semi-definite operators

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

- Enforces non-negativity structurally ($A \succeq 0 \implies g_A \geq 0$) while being linear in $A$. 
Kernel PSD models and sum of squares

\[ g_A(x) = \langle \phi(x), A\phi(x) \rangle, \quad A \in S_+(\mathcal{H}). \]

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 \( W^s_2(\Omega) \).
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 \( W^s_2(\Omega) \).

New problem -

\[
\sup_{c \in \mathbb{R}, \ A \succeq 0} c \quad \text{st} \quad \forall x \in \Omega, \ f(x) - c = \langle \phi(x), A\phi(x) \rangle
\]
Modelling and optimizing \( f \in C^m(\Omega) \): three steps

**Step 1** - Showing the strengthening is reached for some \( k = k_s, s > d/2 \)

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

\[
\sup_{c \in \mathbb{R}} c \quad \text{st} \quad \forall x \in \Omega, \ f(x) - c \geq 0 \tag{3}
\]
Modelling and optimizing $f \in C^m(\Omega)$: three steps

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

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

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

**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 $W^s_2(\Omega)$

\[
f - f_* = \sum_{i=1}^{N} f_i^2, \quad f_i \in W^s_2(\Omega)
\]
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: $\exists A^* \in S_+(H)$ s.t. $f(x) = f_* + \langle k_x, A^* k_x \rangle$

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

$$\hat{c}, \hat{A} = \arg\max_{c \in \mathbb{R}, \ A \in S_+(H)} c - \lambda \text{Tr}(A)$$

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 $k = k_s$, $s > d/2$

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

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

$$\hat{c}, \hat{A} = \arg\max_{c \in \mathbb{R}, A \in S_+(H)} c - \lambda \text{Tr}(A)$$

subject to $f(x_i) - c = \langle k_{x_i}, A k_{x_i} \rangle$, $1 \leq i \leq n$ \hspace{1cm} (4)

- $n$ large enough to guarantee that $\|\hat{c} - f_*\| \leq \epsilon$

- Regularization $\lambda \text{Tr}(A)$ 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 $(x_i)_{1 \leq i \leq n}$:

$$\hat{c}, \hat{A} = \arg\max_{c \in \mathbb{R}, A \in S_+(\mathcal{H})} c - \lambda \text{Tr}(A)$$

subject to $f(x_i) - c = \langle k_{x_i}, A k_{x_i} \rangle$, $1 \leq i \leq n$

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 $\{f - f_* \leq \delta\} \subset \overset{\circ}{\Omega}$ for some $\delta > 0$.

For any $s \in ]d/2, m - 2]$, there exists $h_1, \ldots, h_N \in W^s_2(\Omega)$ such that

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

$$= f_* + \langle k_x, A_* k_x \rangle_{\mathcal{H}} \text{ where } A_* = \sum h_i \otimes h_i$$

- 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 $x_1, \ldots, x_n \in \Omega$ and solve

$$\hat{c}, \hat{A} = \arg\max_{c \in \mathbb{R}, A \succeq 0} c - \lambda \text{Tr}(A) \quad \text{st} \quad f(x_i) = c + \langle \phi(x_i), A\phi(x_i) \rangle.$$ 

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

Up to logarithmic terms: $x_1, \ldots, x_n$ 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 $1 - \delta$:

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

- Near optimal $(\varepsilon^{-d/(m-d/2)}$ for Sobolev).
- In practice, $n$ is a computational budget.
Step 3: Pseudocode for the algorithm

Input: \( f : \mathbb{R}^d \rightarrow \mathbb{R}, \Omega \subset \mathbb{R}^d, n \geq 0, \lambda > 0, s > d/2 \).

1. **Sampling**: \( \{x_1, \ldots, x_n\} \) sampled i.i.d. uniformly on \( \Omega \)

2. **Feature computation**
   - Set \( f_j = f(x_j), \forall j \in \{1, \ldots, n\} \)
   - Compute \( K_{ij} = k_s(x_i, x_j) \)
   - Set \( \Phi_j \in \mathbb{R}^n \) computed using a Cholesky decomposition of \( K \)

3. **Solve**
   \[
   \max_{c \in \mathbb{R}, B \succeq 0} c - \lambda \text{Tr}(B) \quad \text{s. t.} \quad \forall j \in \{1, \ldots, n\}, f_j - c = \Phi_j^\top B \Phi_j
   \]

Output: \( c \) proxy for \( f_\ast \).

Extension to compute \( \hat{x} \) possible
First experiments

Example of function -

Experiments on benchmarks -

<table>
<thead>
<tr>
<th></th>
<th>d</th>
<th>error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trid</td>
<td>6</td>
<td>0.00E+00</td>
</tr>
<tr>
<td>Watson</td>
<td>6</td>
<td>1.09E-03</td>
</tr>
<tr>
<td>Hartmann6</td>
<td>6</td>
<td>0.00E+00</td>
</tr>
<tr>
<td>LennardJones</td>
<td>6</td>
<td>0.00E+00</td>
</tr>
<tr>
<td>Thurber</td>
<td>7</td>
<td>9.70E+03</td>
</tr>
<tr>
<td>Xor</td>
<td>9</td>
<td>6.99E-03</td>
</tr>
<tr>
<td>Paviani</td>
<td>10</td>
<td>1.03E-04</td>
</tr>
<tr>
<td>Cola</td>
<td>17</td>
<td>3.35E-01</td>
</tr>
</tbody>
</table>
Optimization using sum of squares polynomials

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

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

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

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

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

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].
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!