Hard Shape-Constrained Kernel Machines

Pierre-Cyril Aubin-Frankowski[1], Zoltán Szabó[2]

https://arxiv.org/abs/2005.12636

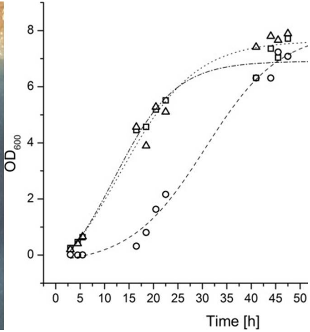
NeurIPS, December 2020

[1]École des Ponts ParisTech, CAS - Centre Automatique et Systèmes, MINES ParisTech

[2]CMAP - Center of Applied Mathematics, CNRS, École Polytechnique, Institut Polytechnique de Paris

What are shape constraints?

Nonparametric estimation



Side information

$\hookrightarrow$ compensates small number of samples or excessive noise

Shape constraints

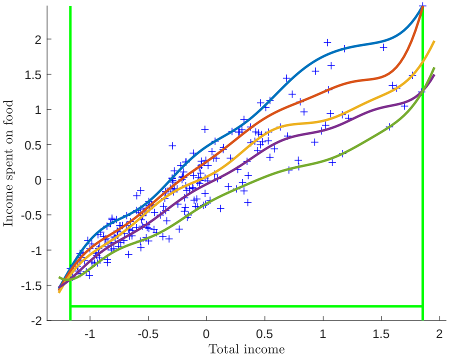
- nonnegative functions
  $f(x) \geq 0$
- monotonic functions
  $f'(x) \geq 0$
- convex functions in 1D
  $f''(x) \geq 0$
- supermodular functions
  $\partial_i \partial_j f(x) \geq 0 \; i \neq j$

Biology, Statistics, Economics, Path-planning, Supply chain,...

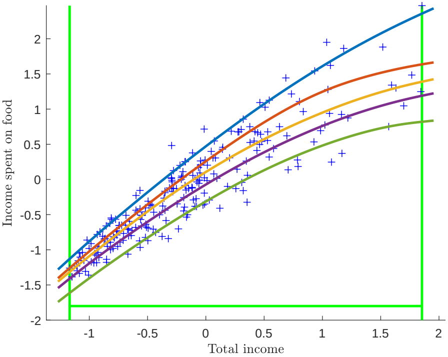
Ubiquitous and handled as a constrained optimization problem

In practice: nonparametric estimation under constraints

In statistics: nonnegative densities, non-crossing quantiles



non-crossing+increasing



non-crossing+increasing+concave

Qualitative priors have a great effect on the shape of solutions!

Problem statement

Given samples $(x_n, y_n)_{n\in[N]} \in (\mathcal{X}\times\mathbb{R})^N$, a *loss* $L : (\mathcal{X}\times\mathbb{R}\times\mathbb{R})^N \to \mathbb{R}\cup\{\infty\}$, a *regularizer* $\Omega : \mathbb{R}_+ \to \mathbb{R}$. Consider

$$\begin{aligned}\bar{f} \in \underset{f\,\in\,\mathcal{F}_k}{\arg\min}\ & \mathcal{L}(f) = L\left((x_n, y_n, f(x_n))_{n\in[N]}\right) + \Omega\left(\|f\|_{\mathcal{F}_k}\right)\\ \text{s.t.}\qquad & b_i \le D_i f(x), \quad \forall\, x\in\mathcal{K}_i,\ \forall i \in [\mathcal{I}] = [\![1,\mathcal{I}]\!].\end{aligned}$$

where $\mathcal{F}_k$ is a RKHS of smooth functions from $\mathcal{X}$ to $\mathbb{R}$, D_i is a differential operator ($D_i = \sum_j \gamma_j \partial^{r_j}$), $b_i \in \mathbb{R}$ is a lower bound, $\mathcal{K}_i$ is compact.

For non-finite $\mathcal{K}_i$, we have an infinite number of constraints!

$\hookrightarrow$ No representer theorem to work in finite dimensions!

How can we make this optimization problem computationally tractable?

Dealing with an infinite number of constraints: an overview

$$\bar{f} \in \arg\min_{f \in \mathcal{F}} \mathcal{L}(f) \text{ s.t. } "b_i \leq D_i f(x), \forall x \in \mathcal{K}_i, \forall i \in [\mathcal{I}]", \mathcal{K}_i \text{ non-finite}$$

Relaxing

- Discretize constraint at "virtual" samples $\{\tilde{x}_{m,i}\}_{m \leq M} \subset \mathcal{K}_i$,
  $\hookrightarrow$ no guarantees out-of-samples [Agrell, 2019, Takeuchi et al., 2006]
- Add constraint-inducing penalty, $\Omega_{\text{cons}}(f) = -\lambda \int_{\mathcal{K}_i} \min(0, D_i f(x) - b_i)\mathrm{d}x$
  $\hookrightarrow$ no guarantees, changes the problem objective [Brault et al., 2019]

Tightening

- Replace $\mathcal{F}$ by algebraic subclass of functions satisfying the constraints
  $\hookrightarrow$ hard to stack constraints, $\Phi(x)^\top A \Phi(x)$ [Marteau-Ferey et al., 2020]
- **Our solution:** discretize $\mathcal{K}_i$ but replace b_i using RKHS geometry
  We show how to tighten an infinite number of affine constraints over a compact set into finitely many SOC constraints in RKHSs
  $\hookrightarrow$ we have a representer theorem!

Reproducing kernel Hilbert spaces (RKHS) at a glance (1)

A RKHS $(\mathcal{F}_k, \langle \cdot, \cdot \rangle_{\mathcal{F}_k})$ is a Hilbert space of real-valued functions over a set $\mathcal{X}$ if one of the following equivalent conditions is satisfied [Aronszajn, 1950]

$\exists\, k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ s.t. $k_x(\cdot) = k(x, \cdot) \in \mathcal{F}_k$ and $f(x) = \langle f(\cdot), k_x(\cdot) \rangle_{\mathcal{F}_k}$

the topology of $(\mathcal{F}_k, \langle \cdot, \cdot \rangle_{\mathcal{F}_k})$ is stronger than pointwise convergence i.e. $\delta_x : f \mapsto f(x)$ is continuous for all x for $f \in \mathcal{F}_k$.

$$|f(x) - f_n(x)| = |\langle f - f_n, k_x \rangle_k| \leq \|f - f_n\|_k \|k_x\|_k = \|f - f_n\|_k \sqrt{k(x, x)}$$

k is s.t. $\exists\, \Phi_k : \mathcal{X} \to \mathcal{F}_k$ s.t. $k(x, y) \; = \; \langle \Phi_k(x), \Phi_k(y) \rangle_{\mathcal{F}_k}$, $\Phi_k(x) = k_x(\cdot)$

k is s.t. $\mathbf{G} = [k(x_i, x_j)]_{i,j=1}^n \succcurlyeq 0$ and $\mathcal{F}_k := \overline{\mathrm{span}(\{k_x(\cdot)\}_{x \in \mathcal{X}})}$, i.e. the completion for the pre-scalar product $\langle k_x(\cdot), k_y(\cdot) \rangle_{k,0} = k(x, y)$

Reproducing kernel Hilbert spaces (RKHS) at a glance (2)

- There is a one-to-one correspondence between kernels k and RKHSs $(\mathcal{F}_k, \langle \cdot, \cdot \rangle_{\mathcal{F}_k})$. Changing $\mathcal{X}$ or $\langle \cdot, \cdot \rangle_{\mathcal{F}_k}$ changes the kernel k.
- for $\mathcal{X} \subset \mathbb{R}^d$, Sobolev spaces $\mathcal{H}^s(\mathcal{X})$ satisfying $s > d/2$ are RKHSs. For $\mathcal{X} = \mathbb{R}^d$ their (Matérn) kernels are well known. Classical kernels include

$$k_{\text{Gauss}}(x, y) = \exp\left(-\|x - y\|^2_{\mathbb{R}^d}/(2\sigma^2)\right) \quad k_{\text{lin}}(x, y) = \langle x, y \rangle_{\mathbb{R}^d}$$

- if $\mathcal{X} \subset \mathbb{R}^d$ is contained in the closure of its interior (e.g. $[0, +\infty[$, for $d = 1$), $k \in \mathcal{C}^{s,s}(\mathcal{X} \times \mathcal{X}, \mathbb{R})$, $D = \sum_j \gamma_j \partial^{\mathbf{r}_j}$ a differential operator of order at most s, then $\mathcal{F}_k \subset \mathcal{C}^s(\mathcal{X}, \mathbb{R})$ and reproducing formula for derivatives:

$$D_x k(x, \cdot) \in \mathcal{F}_k \quad ; \quad Df(x) = \langle f(\cdot), D_x k(x, \cdot) \rangle_{\mathcal{F}_k}$$

Deriving SOC constraints through continuity moduli

Take $\delta \geq 0$ and x s.t. $\|x - \tilde{x}_m\| \leq \delta$

$$
\begin{aligned}
|Df(x) - Df(\tilde{x}_m)| &= |\langle f(\cdot), D_x k(x, \cdot) - D_x k(\tilde{x}_m, \cdot)\rangle_k| \\
&\leq \|f(\cdot)\|_k \underbrace{\sup_{\{x \,|\, \|x-\tilde{x}_m\| \leq \delta\}} \|D_x k(x, \cdot) - D_x k(\tilde{x}_m, \cdot)\|_k}_{\eta_m(\delta)}
\end{aligned}
$$

$$\omega_m(Df, \delta) := \sup_{\{x \,|\, \|x-\tilde{x}_m\| \leq \delta\}} |Df(x) - Df(\tilde{x}_m)| \leq \eta_m(\delta)\|f(\cdot)\|_k$$

For a covering $\mathcal{K} \subset \bigcup_{m\in[M]} \mathbb{B}_{\mathcal{X}}(\tilde{x}_m, \delta_m)$

$$
\begin{aligned}
\text{“}b \leq Df(x),\ \forall x \in \mathcal{K}\text{“} &\Leftarrow \text{“}b + \omega_m(Df, \delta) \leq Df(\tilde{x}_m),\ \forall m \in [M]\text{“} \\
&\Leftarrow \text{“}b + \eta_m\|f(\cdot)\| \leq Df(\tilde{x}_m),\ \forall m \in [M]
\end{aligned}
$$

Since the kernel is smooth, $\delta \to 0$ gives $\eta_m(\delta) \to 0$.

Main theorem

$$(f_\eta, b_\eta) \in \underset{f \in \mathcal{F}_k, b \in \mathcal{B}}{\arg\min} \quad \mathcal{L}(f) = L\left(b, (x_n, y_n, f(x_n))_{n\in[N]}\right) + \Omega\left(\|f\|_k\right)$$

$$\text{s.t.} \quad b_i + \eta_{i,m}\|f(\cdot)\|_k \leq D_i f(\tilde{x}_{m,i}), \quad \forall\, m \in [M_i], \forall i \in [\mathcal{I}].$$

where $\mathcal{B}$ is a closed convex constraint set over $(b_i)_{i\in[\mathcal{I}]}$. If $\Omega(\cdot)$ is strictly increasing, then

Theoretical guarantees [Aubin-Frankowski and Szabó, 2020]

i) The finite number of SOC constraints is **tighter** than the infinite number of affine constraints.

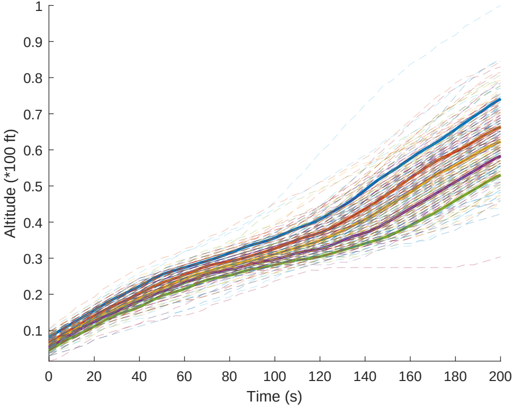
ii) **Representer theorem** (optimal solutions have a finite expression) $f_\eta = \sum_{i\in[\mathcal{I}],m\in[M_i]} \tilde{a}_{i,m} D_{i,x} k\left(\tilde{x}_{i,m}, \cdot\right) + \sum_{n\in[N]} a_n k(x_n, \cdot)$

iii) If $\mathcal{L}$ is μ-strongly convex, we have **bounds**: computable/theoretical

$$\|f_\eta - \bar{f}\|_k \leq \min\left(\sqrt{\frac{2(\mathcal{L}(f_\eta) - \mathcal{L}(f_{\eta=0}))}{\mu}}, \sqrt{\frac{L_{\bar{f}}\|\boldsymbol{\eta}\|_\infty}{\mu}}\right)$$

Joint quantile regression (JQR): airplane data

Airplane trajectories at takeoff have increasing altitude.



JQR with monotonic constraint over $[x_{\min}, x_{\max}]$:

Increasing quantiles should be non-crossing

Data provided by ENAC (flights Paris$\to$Toulouse) [Nicol, 2013]

Two shape constraints jointly handled with $15k$ samples.
Works with higher dimensions too!

Teaser slide

This approach works as well for

- SDP constraints (e.g. convexity for $d \geq 2$): $0 \preccurlyeq \mathbf{Hess}(f)(x)$
- Vector-valued functions $f : \mathcal{X} \to \mathbb{R}^Q$

Control problem: $\mathcal{F}_k$ is a Hilbert space of trajectories $[0, T] \to \mathbb{R}^Q$

$$\begin{aligned} \min_{x(\cdot) \in \mathcal{F}_k} \quad & g(x(T)) + \|x(\cdot)\|_k^2 \\ \text{s.t.} \quad & x(0) = x_0, \\ & c_i(t)^\top x(t) \leq d_i(t), \quad \forall\, t \in [0, T], \forall i \in [\mathcal{I}]. \end{aligned}$$

Stay tuned!

Articles: `https://pcaubin.github.io/`

Code: `https://github.com/PCAubin/Hard-Shape-Constraints-for-Kernels`

References I

Agrell, C. (2019).
Gaussian processes with linear operator inequality constraints.
Journal of Machine Learning Research, 20:1–36.

Aronszajn, N. (1950).
Theory of reproducing kernels.
Transactions of the American Mathematical Society, 68:337–404.

Aubin-Frankowski, P.-C. and Szabó, Z. (2020).
Hard shape-constrained kernel machines.
https://arxiv.org/abs/2005.12636.

Brault, R., Lambert, A., Szabo, Z., Sangnier, M., and d'Alche Buc, F. (2019).
Infinite task learning in RKHSs.
volume 89 of *Proceedings of Machine Learning Research*, pages 1294–1302. PMLR.

Marteau-Ferey, U., Bach, F., and Rudi, A. (2020).
Non-parametric models for non-negative functions.
https://arxiv.org/abs/2007.03926.

References II

Nicol, F. (2013).
Functional principal component analysis of aircraft trajectories.
In *International Conference on Interdisciplinary Science for Innovative Air Traffic Management (ISIATM)*.

Takeuchi, I., Le, Q., Sears, T., and Smola, A. (2006).
Nonparametric quantile estimation.
Journal of Machine Learning Research, 7:1231–1264.