Hard Shape-Constrained Kernel Machines

Pierre-Cyril Aubin-Frankowski\textsuperscript{1}, Zoltán Szabó\textsuperscript{2}


NeurIPS, December 2020

\textsuperscript{1}École des Ponts ParisTech, CAS – Centre Automatique et Systèmes, MINES ParisTech

\textsuperscript{2}CMAP – Center of Applied Mathematics, CNRS, École Polytechnique, Institut Polytechnique de Paris
What are shape constraints?

**Nonparametric estimation**

- Side information 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
  \[ \partial_i \partial_j f(x) \geq 0 \quad i \neq j \]

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

\[
\tilde{f} \in \arg \min_{f \in \mathcal{F}_k} \mathcal{L}(f) = L\left( (x_n, y_n, f(x_n))_{n \in [N]} \right) + \Omega(\|f\|_{\mathcal{F}_k})
\]

s.t. \(b_i \leq D_i f(x), \quad \forall x \in \mathcal{K}_i, \forall i \in [I] = [1, I]\).

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

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

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

Relaxing

- Discretize constraint at “virtual“ samples \( \{\tilde{x}_{m,i}\}_{m \leq M} \subset \mathcal{K}_i \),
  \( \xrightarrow{\text{no guarantees out-of-samples [Agrell, 2019, Takeuchi et al., 2006]}} \)

- Add constraint-inducing penalty, \( \Omega_{\text{cons}}(f) = -\lambda \int_{\mathcal{K}_i} \min(0, D_i f(x) - b_i) dx \)
  \( \xrightarrow{\text{no guarantees, changes the problem objective [Brault et al., 2019]}} \)

Tightening

- Replace \( \mathcal{F} \) by algebraic subclass of functions satisfying the constraints
  \( \xrightarrow{\text{hard to stack constraints, } \Phi(x)\top A \Phi(x) [Marteau-Ferey et al., 2020]} \)

**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
  \( \xrightarrow{\text{we have a representer theorem!}} \)
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]

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

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. \(\exists \Phi_k : \mathcal{X} \rightarrow \mathcal{F}_k\) 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. \(G = [k(x_i, x_j)]_{i,j=1}^n \succeq 0\) and \(\mathcal{F}_k := \text{span}(\{k_x(\cdot)\}_{x \in \mathcal{X}}), \) 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 $X$ or $\langle \cdot, \cdot \rangle_{\mathcal{F}_k}$ changes the kernel $k$.

- for $X \subset \mathbb{R}^d$, Sobolev spaces $\mathcal{H}^s(X)$ satisfying $s > d/2$ are RKHSs. For $X = \mathbb{R}^d$ 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 $X \subset \mathbb{R}^d$ is contained in the closure of its interior (e.g. $[0, +\infty[$, for $d = 1$), $k \in \mathcal{C}^{s,s}(X \times X, \mathbb{R})$, $D = \sum_j \gamma_j \partial^{r_j}$ a differential operator of order at most $s$, then $\mathcal{F}_k \subset \mathcal{C}^s(X, \mathbb{R})$ 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. $\|x - \tilde{x}_m\| \leq \delta$

\[
|Df(x) - Df(\tilde{x}_m)| = |\langle f(\cdot), D_x k(x, \cdot) - D_x k(\tilde{x}_m, \cdot) \rangle_k|
\]

\[
\leq \|f(\cdot)\|_k \sup_{\{x \mid \|x - \tilde{x}_m\| \leq \delta\}} \|D_x k(x, \cdot) - D_x k(\tilde{x}_m, \cdot)\|_k
\]

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

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

"$b \leq Df(x), \forall x \in \mathcal{K}$" $\iff$ "$b + \omega_m(Df, \delta) \leq Df(\tilde{x}_m), \forall m \in [M]$"

\[
\iff "b + \eta_m \|f(\cdot)\| \leq Df(\tilde{x}_m), \forall m \in [M]"
\]

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

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

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

where \(B\) is a closed convex constraint set over \((b_i)_{i \in [I]}\). 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)

\[ f_\eta = \sum_{i \in [I], m \in [M_i]} \tilde{a}_{i,m} D_{i,x} k(\tilde{x}_{i,m}, \cdot) + \sum_{n \in [N]} a_n k(x_n, \cdot) \]

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

\[\|f_\eta - \bar{f}\|_k \leq \min \left( \frac{\sqrt{2(\mathcal{L}(f_\eta) - \mathcal{L}(f_\eta=0))}}{\mu}, \sqrt{\frac{L_{\bar{f}} \|\eta\|_\infty}{\mu}} \right)\]
Joint quantile regression (JQR): airplane data

Airplane trajectories at takeoff have increasing altitude.

JQR with monotonic constraint over \([x_{\text{min}}, x_{\text{max}}]\):

Increasing quantiles should be non-crossing

Data provided by ENAC (flights Paris→Toulouse) [Nicol, 2013]

Two shape constraints jointly handled with 15k samples.
Works with higher dimensions too!
This approach works as well for

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

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

$$\min_{x(\cdot) \in \mathcal{F}_k} g(x(T)) + \|x(\cdot)\|_k^2$$

s.t. $x(0) = x_0,$

$$c_i(t)^\top x(t) \leq d_i(t), \quad \forall t \in [0, T], \forall i \in [I].$$

Stay tuned!

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

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


