Kernels and optimization: Hilbert vs tropical, kernel Sum-of-Squares, optimal control, c-concavity and representer theorems

Pierre-Cyril Aubin-Frankowski

INRIA SIERRA, Paris, France
Moving to TU Wien for postdoc in Sept 23

LIKE 2023

June 29, 2023, Bern

A very natural problem

Let X be a set, and $\mathcal{F} = \{f : X \to \mathbb{R}\}$ a function class. For $F \in \mathcal{F}$ and $L : \mathcal{F} \to \mathbb{R}$

$$\min_{x \in X} F(x) \quad \text{VS} \quad \min_{f \in \mathcal{F}} \mathcal{L}(f) = L(f(x_1), \ldots, f(x_N))$$

Typical examples of $\mathcal{F}$ in this talk

- $\mathcal{F}$ is a RKHS $\mathcal{H}_k$ with kernel k
- $\mathcal{F}$ is CVEX($\mathbb{R}^d$), the set of convex lower semicontinuous functions over $\mathbb{R}^d$
- $\mathcal{F}$ is Lip(X), the set of 1-Lipschitz functions over a metric space X

Questions:

- can we minimize a given F through function evaluations?
- can we minimize over $\mathcal{F}$ when $\mathcal{L}$ involves a finite number of evaluations?

Some very special function spaces, the ones generated by a kernel

RKHSs and convex functions have the common property of having clear generators:

$$\mathcal{H}_k = \{f(\cdot) = \Sigma_{y\in X} a_y k(\cdot, y) \,|\, (a_y)_y \text{ finite}\} \textbf{ + completion}$$

$$\mathrm{CVEX}(\mathbb{R}^d) = \{f(\cdot) = \sup_{y\in\mathbb{R}^d} (\cdot, y) + a_y \,|\, (a_y)_y \subset \mathbb{R}\cup\{-\infty\}\}$$

Some very special function spaces, the ones generated by a kernel

RKHSs and convex functions have the common property of having clear generators:

$$\mathcal{H}_k = \{f(\cdot) = \textstyle\sum_{y\in X} a_y k(\cdot, y) \,|\, (a_y)_y \text{ finite}\} + \textbf{completion}$$

$$\mathrm{CVEX}(\mathbb{R}^d) = \{f(\cdot) = \sup_{y\in\mathbb{R}^d} (\cdot, y) + a_y \,|\, (a_y)_y \subset \mathbb{R}\cup\{-\infty\}\}$$

More generally take a (max-plus) kernel $b : X \times Y \to \mathbb{R}$, and define its *range*

$$\mathrm{Rg}(B) := \{\sup_{y\in Y} b(\cdot, y) + a_y \,|\, a_y \in \mathbb{R}\cup\{-\infty\}\}$$

Take $X = Y$ for now:

i) For $X = \mathbb{R}^d$, $b(x, y) = -\|x - y\|^2$ gives the 1-semiconvex l.s.c. functions,

$$\mathrm{Rg}(B) = \{f \text{ l.s.c.} \,|\, f + \|\cdot\|^2 \text{ is convex}\}.$$

ii) For (X, d) a metric space, $p \in (0, 1]$, $b(x, y) = -d(x, y)^p$ gives the $(1, p)$-Hölder continuous functions,

$$\mathrm{Rg}(B) = \{f \,|\, \forall x, y\,, |f(x) - f(y)| \leq 1 \cdot d(x, y)^p\}.$$

What are we going to see?

If $\mathcal{F} = \mathcal{H}_k$ is a RKHS,

- (minimize over $\mathcal{H}_k$): **known** $\to$ representer theorems
  $\hookrightarrow$ (**new** cases in optimal control/estimation)
- (minimize $F \in \mathcal{H}_k$): **new** $\to$ kernel Sum-of-Squares

If $\mathcal{F} = \mathrm{Rg}(B)$ is a tropical kernel space,

- (minimize over $\mathrm{Rg}(B)$): **new** $\to$ tropical representer theorems
- (minimize $F \in \mathrm{Rg}(B)$): **new** $\to$ F c-concave and alternating minimization

Separate works with Alain Bensoussan (UT Dallas), Alessandro Rudi (INRIA Paris), Stéphane Gaubert (INRIA Polytechnique), Flavien Léger (INRIA Paris)

Optimizing over RKHSs: representer theorem

Typical representer theorem e.g. B. Schölkopf, R. Herbrich, and A. J. Smola. "A Generalized Representer Theorem". In: *Computational Learning Theory (CoLT)*. 2001, pp. 416–426

Let $L : \mathbb{R}^N \to \mathbb{R} \cup \{\infty\}$, strictly increasing $\Omega : \mathbb{R}_+ \to \mathbb{R}$, and assume there exists

$$\bar{f} \in \operatorname{argmin}_{f \in \mathcal{H}_k} L\left((f(x_n))_{n\in[N]}\right) + \Omega\left(\|f\|_k\right)$$

Then $\exists\, (a_n)_{n\in[N]} \in \mathbb{R}^N$ s.t. $\bar{f}(\cdot) = \sum_{n\in[N]} a_n k(\cdot, x_n)$

$\hookrightarrow$ Actually even for $\Omega = 0$, existence of $\bar{f}$, gives existence of optimal $\bar{f}_0(\cdot) = \sum_{n\in[N]} a_n k(\cdot, x_n)$.

$\hookrightarrow$ All vs some optimal solutions lie in a finite dimensional subspace of $\mathcal{H}_k$.

Finite number of evaluations $\implies$ finite number of coefficients

What if there is no RKHS? Find one! Example in optimal control

The Linear-Quadratic (LQ) optimal control is defined over

$$\mathcal{S}_{[t_0,T]} := \{x(\cdot) \,|\, x(t_0) = 0, \exists\, u(\cdot) \in L^2(t_0, T) \text{ s.t. } x'(t) = Ax(t) + Bu(t) \text{ a.e. }\}$$

a vector space of controlled trajectories $x(\cdot) : [t_0, T] \to \mathbb{R}^Q$.

LQ optimal control

$$\min_{x(\cdot)\in\mathcal{S}_{[t_0,T]}u(\cdot)\in L^2} g(x(T)) + \int_{t_0}^{T} \|u(\tau)\|^2 \mathrm{d}\tau$$

with $u(t) = B^{\ominus}[x'(t) - Ax(t)]$

What if there is no RKHS? Find one! Example in optimal control

The Linear-Quadratic (LQ) optimal control is defined over

$$\mathcal{S}_{[t_0,T]} := \{x(\cdot) \,|\, x(t_0) = 0,\ \exists\, u(\cdot) \in L^2(t_0, T) \text{ s.t. } x'(t) = Ax(t) + Bu(t) \text{ a.e. }\}$$

a vector space of controlled trajectories $x(\cdot) : [t_0, T] \to \mathbb{R}^Q$.

LQ optimal control

$$\min_{x(\cdot)\in\mathcal{S}_{[t_0,T]} u(\cdot)\in L^2} g(x(T)) + \int_{t_0}^{T} \|u(\tau)\|^2 \mathrm{d}\tau$$

with $u(t) = B^{\ominus}[x'(t) - Ax(t)]$

"KRR" (Kernel Ridge Regression)

$$\min_{x(\cdot)\in\mathcal{S}_{[t_0,T]}} g(x(T)) + \|x(\cdot)\|^2_{\mathcal{S}_{[t_0,T]}}$$

with $\|x(\cdot)\|^2_{\mathcal{S}_{[t_0,T]}} = \|\mathbf{B}^{\ominus}[x'(\cdot) - Ax(\cdot)]\|^2_{L^2(t_0,T)}$

The corresponding kernel has the form of a Gramian:

$$K(s,t) = \int_{t_0}^{\min(s,t)} e^{A(s-\tau)} B(\tau) B(\tau)^\top e^{A^\top(t-\tau)} \mathrm{d}\tau.$$

and the optimal solution is of the form $\bar{x}(\cdot) = K(\cdot, T)p_T$ for some $p_T \in \mathbb{R}^Q$.

#1 Where's ~~Waldo/Charlie~~ the kernel? For Kalman estimation

Continuous-time estimation problem (smoothing/filtering) over GPs with linear SDE

$$dx(t) = Fx(t)dt + Gdw(t), \qquad x(t_0) = \xi, \tag{1}$$

$$dy(t) = Hx(t)dt + db(t), \qquad y(t_0) = 0. \tag{2}$$

Problem: Estimate $x(s)$ with the σ-algebra $\mathcal{Y}^T = \sigma(y(\tau), 0 \leq \tau \leq T)$ by (linear) minimum mean square estimator, a.k.a. the minimum variance linear estimator

$$\hat{x}(s|T) = \mathbb{E}[x(s)|\mathcal{Y}^T] = x_S(s|T) := \bar{x}(s) + \int_{t_0}^{T} S_s(t|T)dy(t). \tag{3}$$

#1 Where's ~~Waldo/Charlie~~ the kernel? For Kalman estimation

Continuous-time estimation problem (smoothing/filtering) over GPs with linear SDE

$$dx(t) = Fx(t)dt + Gdw(t), \qquad x(t_0) = \xi, \tag{1}$$

$$dy(t) = Hx(t)dt + db(t), \qquad y(t_0) = 0. \tag{2}$$

Problem: Estimate $x(s)$ with the σ-algebra $\mathcal{Y}^T = \sigma(y(\tau), 0 \leq \tau \leq T)$ by (linear) minimum mean square estimator, a.k.a. the minimum variance linear estimator

$$\hat{x}(s|T) = \mathbb{E}[x(s)|\mathcal{Y}^T] = x_S(s|T) := \bar{x}(s) + \int_{t_0}^{T} S_s(t|T)dy(t). \tag{3}$$

$$\epsilon_S(s|T) := x(s) - x_S(s|T) = x(s) - \int_{t_0}^{T} S_s(t|T)dy(t). \tag{4}$$

$$\hat{S}_s(\cdot|T) \in \operatorname{argmin}_{S(\cdot|T)} \Gamma_S(s|T) = \mathbb{E}[\epsilon_S(s|T)(\epsilon_S(s|T))^*]. \tag{5}$$

The kernel is the covariance of $\epsilon_{\hat{S}_s}(\cdot|T)$ and we have $\hat{S}_s(t|T) = K(s,t|T)H^*R^{-1}$,

$$K(s,t|T) = \mathbb{E}[\epsilon_{\hat{S}_s}(s|T)(\epsilon_{\hat{S}_t}(t|T))^*] \in \mathcal{L}(\mathbb{R}^{n,*}, \mathbb{R}^n) \tag{6}$$

#2 Where's ~~Waldo/Charlie~~ the kernel? For least squares estimation

Using least squares formulation of the estimation problem

$$L_x(x(\cdot)) := \int_{t_0}^{T} \|y(t) - Hx(t)\|^2_{R^{-1}} dt + \|G^{\ominus}(x'(t) - Fx(t))\|^2_{Q^{\ominus}} dt + \langle \Pi_0^{\ominus} x(t_0), x(t_0)\rangle + \langle \Sigma_T x(T), x(T)\rangle$$

Introduce the RKHS $\mathcal{S}_{[t_0,T]} = \{x(\cdot) \in H^1 \mid \exists\, u(\cdot) \in L^2 \text{ s.t. } x'(\tau) = Fx(\tau) + GQ^{\frac{1}{2}} u(\tau)\}$.

$$\|x(\cdot)\|^2_{\mathcal{S}_{[t_0,T]}} = \langle \Pi_0^{-1} x(t_0), x(t_0)\rangle + \langle \Sigma_T x(T), x(T)\rangle + \int_{t_0}^{T} \|u(\tau)\|^2 d\tau + \int_{t_0}^{T} \langle H^* R^{-1} H x(\tau), x(\tau)\rangle\, d\tau$$

#2 Where's ~~Waldo/Charlie~~ the kernel? For least squares estimation

Using least squares formulation of the estimation problem

$$L_x(x(\cdot)) := \int_{t_0}^{T} \|y(t) - Hx(t)\|^2_{R^{-1}} dt + \|G^{\ominus}(x'(t) - Fx(t))\|^2_{Q^{\ominus}} dt + \langle \Pi_0^{\ominus} x(t_0), x(t_0)\rangle + \langle \Sigma_T x(T), x(T)\rangle$$

Introduce the RKHS $\mathcal{S}_{[t_0,T]} = \{x(\cdot) \in H^1 \mid \exists\, u(\cdot) \in L^2 \text{ s.t. } x'(\tau) = Fx(\tau) + GQ^{\frac{1}{2}}u(\tau)\}$.

$$\|x(\cdot)\|^2_{\mathcal{S}_{[t_0,T]}} = \langle \Pi_0^{-1} x(t_0), x(t_0)\rangle + \langle \Sigma_T x(T), x(T)\rangle + \int_{t_0}^{T} \|u(\tau)\|^2 d\tau + \int_{t_0}^{T} \langle H^* R^{-1} Hx(\tau), x(\tau)\rangle\, d\tau$$

Taking Fréchet derivative (rather than representer theorem)

$$\int_{t_0}^{T} K(\cdot, t|T) H^* R^{-1} y(t) dt = \operatorname{argmin}_{x(\cdot)\in\mathcal{S}} \|R^{-1/2} y(\cdot)\|^2_{L^2} + \|x(\cdot)\|^2_{\mathcal{S}} - 2\left\langle H^*(\cdot) R^{-1}(\cdot) y(\cdot), x(\cdot)\right\rangle_{L^2([t_0,T])}$$

and the kernel has the explicit form (based on Riccati matrices and some semi-groups)

$$K(s,t|T) = \Phi_{F,\Sigma}(s,t_0)(\Pi_0^{-1} + \Sigma(t_0))^{-1}\Phi^*_{F,\Sigma}(t,t_0) + \int_{t_0}^{\min(s,t)} \Phi_{F,\Sigma}(s,\tau) GQG^* \Phi^*_{F,\Sigma}(t,\tau) d\tau \quad (7)$$

What if there is no RKHS? Find one!

- finding an RKHS somewhere allows for simpler computations (representer theorems + kernel trick)
- in LQ optimal control, RKHSs come from vector spaces of trajectories[1]

LQ optimal control $\subset$ kernel methods

- in linear estimation, kernels come from covariances of optimal errors[2]

New formulas for the covariances of GPs induced by linear SDEs!

Now back to minimizing functions rather than over functions.

[1] Pierre-Cyril Aubin-Frankowski. "Linearly Constrained Linear Quadratic Regulator from the Viewpoint of Kernel Methods". In: *SIAM Journal on Control and Optimization* 59.4 (2021), pp. 2693–2716.

[2] Pierre-Cyril Aubin-Frankowski and Alain Bensoussan. "The reproducing kernel Hilbert spaces underlying linear SDE Estimation, Kalman filtering and their relation to optimal control". In: *Pure and Applied Functional Analysis* (2022).

Optimizing a smooth function ~~in a RKHS~~: kernel Sum-of-Squares

Take $F \in \mathcal{H}_k$ with $k \in C^{s_k}(X \times X, \mathbb{R})$, $s_k \geq 0$, $X \subset \mathbb{R}^d$ bounded open. Global optimization of

$$\min_{x \in X} F(x)$$

is in general <u>**non-convex**</u>. BUT it can be rewritten as

$$\sup_{\substack{c \in \mathbb{R} \\ F(x) - c \geq 0, \forall x \in X}} c$$

This convex problem has an infinite number of affine constraints. . . Lets sample them!

Optimizing a smooth function ~~in a RKHS~~: kernel Sum-of-Squares

Take $F \in \mathcal{H}_k$ with $k \in C^{s_k}(X \times X, \mathbb{R})$, $s_k \geq 0$, $X \subset \mathbb{R}^d$ bounded open. Global optimization of

$$\min_{x \in X} F(x)$$

is in general **non-convex**. BUT it can be rewritten as

$$\sup_{\substack{c \in \mathbb{R} \\ F(x) - c \geq 0, \forall x \in X}} c$$

This convex problem has an infinite number of affine constraints... Lets sample them! However, we would get $\hat{c} = \min_{m \in [M]} F(x_m)$ and in the worst case

$$|\hat{c} - \min F| \propto \mathrm{Lip}(F) \cdot h_M \quad \text{where} \quad h_M = \sup_{x \in X} \min_{m \in [M]} \|x - x_m\| \text{ (fill distance)} \tag{8}$$

BUT $h_M \propto \frac{1}{M^d} \to$ **curse of dimensionality**. Can we do better by leveraging the smoothness?

Optimizing a smooth function ~~in a RKHS~~: kernel Sum-of-Squares

We want to do global zero-th order optimization of smooth functions. Scattering inequalities tell us that if $f(x_m) - g(x_m) = 0$ with $f, g \in C^s$, then on a small neighborhood of size r

$$|f(x) - g(x)| \leq C \cdot r^s$$

Question: Can we find a "nice" function $g(x) \geq 0$, $g \in C^2$ such that

$$\sup_{\substack{c \in \mathbb{R} \\ F(x) - c = g(x), \forall x \in X}} c$$

Yes. . . but that's not trivial because of the nonnegativity constraint.

Optimizing a smooth function ~~in a RKHS~~: kernel Sum-of-Squares

We want to do global zero-th order optimization of smooth functions. Scattering inequalities tell us that if $f(x_m) - g(x_m) = 0$ with $f, g \in C^s$, then on a small neighborhood of size r

$$|f(x) - g(x)| \leq C \cdot r^s$$

Question: Can we find a "nice" function $g(x) \geq 0$, $g \in C^2$ such that

$$\sup_{\substack{c \in \mathbb{R} \\ F(x) - c = g(x), \forall x \in X}} c$$

Yes... but that's not trivial because of the nonnegativity constraint.
Can we set $g = h^2$ for some function h? Yes, if $F \in C^2$ has a strictly positive Hessian at a unique global minimum. BUT we don't know how to compute it.

Can we look for h in a RKHS? Yes but non convex equality constraint...

A nice class of nonnegative functions: kernel Sum-of-Squares/PSD models

How to build a nonnegative function given an embedding $\phi : X \to \mathcal{H}_\phi$? Square it!

$$f : x \mapsto \langle \phi(x), \phi(x) \rangle_{\mathcal{H}_\phi} = k_\phi(x, x) \geq 0$$

More generally take a positive semidefinite operator $A \in S^+(\mathcal{H}_\phi)$,

$$f_A : x \mapsto \langle \phi(x), A\phi(x) \rangle_{\mathcal{H}_\phi} \geq 0$$

$$\text{(PSD model)} \quad A = \sum_{i,j=1}^{N} a_{ij} \phi(x_i) \otimes \phi(x_j) \implies f_A(x) = \sum_{i,j=1}^{N} a_{ij} k_\phi(x, x_i) k_\phi(x, x_j)$$

$$\text{(kernel SoS)} \quad [a_{ij}]_{i,j} = \sum_i u_i u_i^\top \text{ (SVD)} \implies f_A(x) = \sum_{i=1}^{N} \Big(\sum_{j=1}^{N} u_{i,j} k_\phi(x, x_j)\Big)^2$$

Note that in general $f_A \notin \mathcal{H}_\phi$ but $f_A \in \mathcal{H}_\phi \odot \mathcal{H}_\phi$ (Hadamard product). If $\operatorname{span}(\{k_\phi(\cdot, x)\}_{x \in X})$ is dense in continuous functions, so are the $\{f_A\}_{A \in S^+(\mathcal{H}_\phi)}$ in nonnegative functions.

Optimization with kernel Sum-of-Squares/PSD models

We can consider the convex problem and approximate it through sampling+regularization[3]

$$\sup_{\substack{c\in\mathbb{R},\, A\in S^+(\mathcal{H}_\phi)\\ F(x)-c=\langle\phi(x),A\phi(x)\rangle_{\mathcal{H}_\phi},\forall x\in X}} c \quad\longrightarrow\quad \sup_{\substack{c\in\mathbb{R},\, A\in S^+(\mathcal{H}_\phi)\\ F(x_m)-c=\langle\phi(x_m),A\phi(x_m)\rangle_{\mathcal{H}_\phi},\forall m\in[M]}} c-\lambda\,\mathrm{Tr}(A)$$

We do have a representer theorem! Two cases[a] for $F \in C^s$:

- if $\exists A^* \in S^+(\mathcal{H}_\phi)$, $F(x) - \min F = \langle\phi(x), A^*\phi(x)\rangle_{\mathcal{H}_\phi}$ then $|\hat{c} - \min F| \leq C_0(F)\cdot h_M^s \propto \frac{1}{M^{\frac{d}{s}}}$
- otherwise, $|\hat{c} - \min F| \leq C_0(F)\cdot h_M \propto \frac{1}{M^d}$.

[a] Pierre-Cyril Aubin-Frankowski and Alessandro Rudi. "Approximation of optimization problems with constraints through kernel Sum-Of-Squares". In: (2022). https://arxiv.org/abs/2301.06339.

Now back to minimizing over functions rather than functions.

[3] Alessandro Rudi, Ulysse Marteau-Ferey, and Francis Bach. *Finding Global Minima via Kernel Approximations*. 2020. arXiv: 2012.11978 [math.OC].

Optimization on tropical function spaces

Take a (max-plus) kernel $b : X \times Y \to \mathbb{R}$, and recall what is the *range*

$$\mathrm{Rg}(B) := \{\sup_{y \in Y} b(\cdot, y) + a_y \mid a_y \in \mathbb{R} \cup \{-\infty\}\}.$$

Given a subset $\hat{X} = \{x_m\}_{m \in \mathcal{I}}$, define

$$\mathrm{Rg}_{\partial\text{-}\hat{X}}(B) := \Big\{ f \in \mathrm{Rg}(B) \,|\, \forall\, m \in \mathcal{I},\, \exists\, p_m \in Y \text{ maximizing: } f(x_m) = \sup_{p \in Y} b(x_m, p) - \sup_{x' \in X} (b(x', p) - f(x')) \Big\}.$$

Optimization on tropical function spaces

Take a (max-plus) kernel $b : X \times Y \to \mathbb{R}$, and recall what is the *range*

$$\mathrm{Rg}(B) := \{\sup_{y \in Y} b(\cdot, y) + a_y \mid a_y \in \mathbb{R} \cup \{-\infty\}\}.$$

Given a subset $\hat{X} = \{x_m\}_{m \in \mathcal{I}}$, define

$$\mathrm{Rg}_{\partial\text{-}\hat{X}}(B) := \Big\{ f \in \mathrm{Rg}(B) \mid \forall\, m \in \mathcal{I}, \exists\, p_m \in Y \text{ maximizing:} \quad f(x_m) = \sup_{p \in Y} b(x_m, p) - \sup_{x' \in X} \big(b(x', p) - f(x')\big) \Big\}.$$

When $b = \langle \cdot, \cdot \rangle$, each p_m can be interpreted as a subgradient at x_m. There is a well-known property in convex regression, (Boyd and Vandenberghe, *Convex Optimization*[Section 6.5.5])

$$\min_{f \in \mathrm{CVEX}} \sum |f(x_m) - \bar{y}_m|^2 \quad \Leftrightarrow \quad \min_{\substack{(p_m, y_m)_{m \in \mathcal{I}} \in (\mathbb{R}^d \times \mathbb{R})^M, \\ y_n - y_m \geq (x_n, p_m)_2 - (x_m, p_m)_2}} \sum |y_m - \bar{y}_m|^2.$$

Question: Can we do the same for more general tropical kernels b?

Optimization on tropical function spaces: interpolation theorem

Proposition (Tropical interpolation)

Let $\mathcal{I}$ be a nonempty index set, given $(x_m, y_m)_{m\in\mathcal{I}} \in (X \times \mathbb{R})^{\mathcal{I}}$, setting $\hat{X} = \{x_m\}_{m\in\mathcal{I}}$, the three following statements are equivalent:

i) *there exists $f \in \mathrm{Rg}_{\partial\text{-}\hat{X}}(B)$ such that $y_m = f(x_m)$ for all $m \in \mathcal{I}$;*

ii) *there exists $(p_m)_{m\in\mathcal{I}} \in (Y)^{\mathcal{I}}$ such that $y_m = f^0(x_m)$ for all $m \in \mathcal{I}$, for*

$$f^0(\cdot) := \max_{m\in\mathcal{I}} b(\cdot, p_m) - b(x_m, p_m) + y_m;$$

iii) *there exists $(p_m)_{m\in\mathcal{I}} \in (Y)^{\mathcal{I}}$ such that $y_n - y_m \geq b(x_n, p_m) - b(x_m, p_m)$ for all $n, m \in \mathcal{I}$.*

Optimization on tropical function spaces: representer theorem

Corollary (Representer theorem)

Given points $(x_m)_{m\in\mathcal{I}} \in X^{\mathcal{I}}$ *and a function* $\mathcal{L} : \mathbb{R}^{\mathcal{I}} \to \mathbb{R}$*, fix* $\hat{X} = \{x_m\}_{m\in\mathcal{I}}$*. Then, if the problem*

$$\min_{f\in\mathrm{Rg}(B)} \mathcal{L}((f(x_m))_{m\in\mathcal{I}}) \tag{9}$$

has a solution $\bar{f} \in \mathrm{Rg}_{\partial\text{-}\hat{X}}(B)$ *with finite values* $(f(x_m))_{m\in\mathcal{I}} \in \mathbb{R}^{\mathcal{I}}$*, it also has a solution* f^0 *as in Proposition 1-ii) which can be obtained solving*

$$\min_{(p_m,y_m)_{m\in\mathcal{I}}\in(Y\times\mathbb{R})^M} \mathcal{L}((y_m))_{m\in\mathcal{I}}) \tag{10}$$

$$\text{s.t. } y_n - y_m \geq b(x_n, p_m) - b(x_m, p_m),\ \forall\, n, m \in \mathcal{I}.$$

Conversely, if (10) *has a solution, then it is also a solution in* $\mathrm{Rg}_{\partial\text{-}\hat{X}}(B)$ *of* (9).

WE DO NOT NEED ANY PROPERTY OF THE KERNEL *b*!

Recall Aronszajn's theorem

Theorem

Given a kernel $k : X \times X \to \mathbb{R}$*, the three following properties are equivalent:*

i) k *is a positive semidefinite kernel, i.e. a kernel being both:*
- *symmetric:* $\forall\, x, y \in X,\ k(x,y) = k(y,x)$*, and*
- *positive:* $\forall\, M \in \mathbb{N}^*, \forall\, (a_m, x_m) \in (\mathbb{R} \times X)^M, \sum_{n,m=1}^{M} a_n a_m k(x_n, x_m) \geq 0$*;*

ii) there exists a Hilbert space $(\mathcal{H}, (\cdot,\cdot)_{\mathcal{H}})$ *and a feature map* $\Phi : X \to \mathcal{H}$ *such that*
- $\forall\, x, y \in X,\ k(x,y) = (\Phi(x), \Phi(y))_{\mathcal{H}}$*;*

iii) k *is the reproducing kernel of the Hilbert space (RKHS) of functions* $\mathcal{H}_k := \overline{\mathcal{H}_{k,0}}$*, the completion for the pre-scalar product* $(k(\cdot,x), k(\cdot,y))_{k,0} = k(x,y)$ *of the space* $\mathcal{H}_{k,0} := \mathrm{span}(\{k(\cdot,x)\}_{x \in X})$*, in the sense that*
- $\forall\, x \in X,\ k(\cdot,x) \in \mathcal{H}_k$ *and* $\forall\, f \in \mathcal{H},\ f(x) = (f, k(\cdot,x))_{\mathcal{H}}$*.*

Main (informal) theorem: Aronszajn's analogue

Theorem (Tropical analogue of Aronszajn theorem)

Given a kernel $b : X \times X \to \mathbb{R} \cup \{-\infty\}$, *the three following properties are equivalent*

i) *b is a tropically positive semidefinite kernel, i.e. symmetric and* $b(x,x) + b(y,y) \geq b(x,y) + b(y,x)$;

ii) *there exists a factorization of b by a feature map* $\psi : X \to \mathbb{R}_{\max}^{\mathcal{Z}}$ *for some set* $\mathcal{Z}$, $b(x,y) = \sup_{z \in \mathcal{Z}} \psi(x,z) + \psi(y,z)$;

iii) *b is the sesquilinear reproducing kernel of a max-plus space of functions* $\mathrm{Rg}(B)$, *the max-plus completion of* $\{\sup_{n \in \{1,\ldots,N\}} a_n + b(\cdot, x_n) \,|\, N \in \mathbb{N}^*, a_n \in \mathbb{R}, x_n \in X\}$, *and b defines a tropical Cauchy-Schwarz inequality over* $\mathbb{R}^X$.

Some kernels b exhibit analogue properties to RKHSs! Are they useful? TBC

Full analogy between Hilbertian and tropical kernels

Dedicated to kernel lovers:[4]

Concept	Hilbertian kernel	Tropical kernel
symmetry	$k(x,y)=k(y,x)$	$b(x,y)=b(y,x)$
positivity	$\sum_{i,j} a_i a_j k(x_i,x_j)\geq 0$	$b(x,x)+b(y,y)\geq b(x,y)+b(y,x)$
feature map	$k(x,y)=(\Phi(x),\Phi(y))_{\mathcal{H}}$	$b(x,y)=\sup_{z\in\mathcal{Z}}\psi(x,z)+\psi(y,z)$
duality bracket	$\langle\mu,f\rangle_{\mathbb{R}^{X,*}\times\mathbb{R}^X}=\int_X f(y)\mathrm{d}\mu(y)$	$\langle\hat{g},f\rangle=\sup_{x\in X} f(x)-\hat{g}(x)$
kernel operator	$K(\mu)(x)=\int_X k(x,y)\mathrm{d}\mu(y)$	$\bar{B}(\hat{f})(x)=\sup_{y\in X} b(x,y)-\hat{f}(y)$
monotone operator	$\langle\mu,K(\mu)\rangle_{\mathbb{R}^{X,*}\times\mathbb{R}^X}\geq 0$	$\langle\hat{f},\bar{B}\hat{f}\rangle+\langle\hat{g},\bar{B}\hat{g}\rangle\geq\langle\hat{f},\bar{B}\hat{g}\rangle+\langle\hat{g},\bar{B}\hat{f}\rangle$
function space	$\mathcal{H}_k=\overline{\mathrm{span}(\{k(\cdot,x)\}_{x\in X})}$	$\mathrm{Rg}(B)=\{\sup_{x\in X}[a_x+b(\cdot,x)]\mid a_x\in\mathbb{R}\}$
reproducing property	$f(x)=(k(\cdot,x),f(\cdot))_{\mathcal{H}_k}$	$\hat{g}(x)=\langle\bar{B}\hat{g},\bar{B}\delta_x^\top\rangle=(\bar{B}\hat{g})(x)$

Now back to minimizing functions rather than over functions.

[4] Pierre-Cyril Aubin-Frankowski and Stéphane Gaubert. "Tropical reproducing kernels and optimization". In: *Integral Equations and Operator Theory* (2023). (to be published).

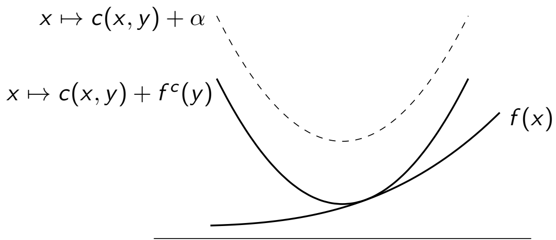
c-concavity

Definition (c-concavity)

We say that a function $f\colon X \to \mathbb{R}$ is c-concave if there exists a function $h\colon Y \to \mathbb{R}$ such that

$$f(x) = \inf_{y\in Y} c(x,y) + h(y), \tag{11}$$

for all $x \in X$. If f is c-concave, then we can take $h(y) = f^c(y) = \sup_{x'\in X} f(x') - c(x',y)$.



NB: Costs c are the opposite of the tropical kernels b (sign convention problem).

For $c = \frac{L}{2}\|x-y\|^2$, c-concave $\Leftrightarrow \nabla^2 f \leq L$.

Majorization–minimization

Let $f : X \to \mathbb{R}$ where X is any set. Choose another set Y and a function $c(x, y)$. Define the upperbound

$$f(x) \le \phi(x, y) := c(x, y) + f^c(y) := c(x, y) + \sup_{x' \in X} f(x') - c(x', y) \tag{12}$$

Do alternating minimization (AM) of the surrogate

$$y_{n+1} = \mathrm{argmin}_{y \in Y}\, c(x_n, y) + f^c(y), \tag{13}$$

$$x_{n+1} = \mathrm{argmin}_{x \in X}\, c(x, y_{n+1}) + f^c(y_{n+1}). \tag{14}$$

Majorization–minimization

Let $f: X \to \mathbb{R}$ where X is any set. Choose another set Y and a function $c(x, y)$. Define the upperbound

$$f(x) \leq \phi(x, y) := c(x, y) + f^c(y) := c(x, y) + \sup_{x' \in X} f(x') - c(x', y) \tag{12}$$

Do alternating minimization (AM) of the surrogate

$$y_{n+1} = \text{argmin}_{y \in Y}\, c(x_n, y) + f^c(y), \tag{13}$$

$$x_{n+1} = \text{argmin}_{x \in X}\, c(x, y_{n+1}) + f^c(y_{n+1}). \tag{14}$$

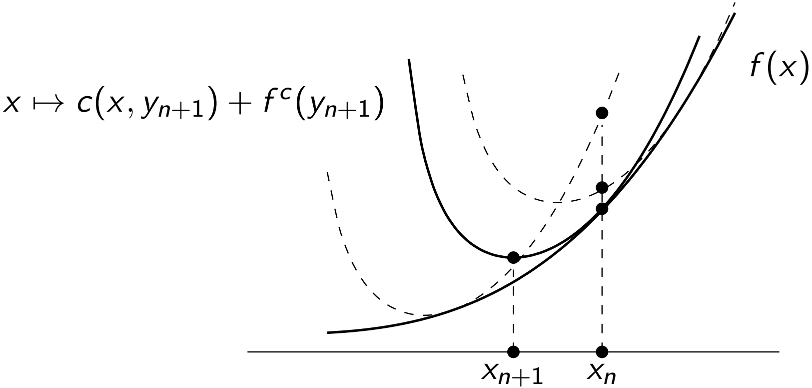
If we can differentiate and $f(x) = \inf_y c(x, y) + f^c(y)$ (c-concavity) then we can write (applying the envelope theorem $\nabla f(x) = \nabla_1 \phi(x, \bar{y}(x))$)

$$-\nabla_x c(x_n, y_{n+1}) = -\nabla f(x_n), \tag{15}$$

$$\nabla_x c(x_{n+1}, y_{n+1}) = 0. \tag{16}$$

Sketch of alternating minimization

$$y_{n+1} = \text{argmin}_{y \in Y}\, c(x_n, y) + f^c(y),$$
$$x_{n+1} = \text{argmin}_{x \in X}\, c(x, y_{n+1}) + f^c(y_{n+1}).$$



Gradient descent with a general cost - Examples

$$-\nabla_x c(x_n, y_{n+1}) = -\nabla f(x_n),$$
$$\nabla_x c(x_{n+1}, y_{n+1}) = 0.$$

In the following: $Y = X$, and c is minimal on the diagonal $\{x = y\}$, so $x_{n+1} = y_{n+1}$

i) Gradient descent: $c(x, y) = \frac{L}{2}\|x - y\|^2$ and $x_{n+1} - x_n = -\frac{1}{L}\nabla f(x_n)$.

ii) Mirror descent: $c(x, y) = u(x|y)$, so $\nabla u(x_{n+1}) - \nabla u(x_n) = -\nabla f(x_n)$.

Gradient descent with a general cost - Examples

$$-\nabla_x c(x_n, y_{n+1}) = -\nabla f(x_n),$$
$$\nabla_x c(x_{n+1}, y_{n+1}) = 0.$$

In the following: $Y = X$, and c is minimal on the diagonal $\{x = y\}$, so $x_{n+1} = y_{n+1}$

i) Gradient descent: $c(x, y) = \frac{L}{2}\|x - y\|^2$ and $x_{n+1} - x_n = -\frac{1}{L}\nabla f(x_n)$.

ii) Mirror descent: $c(x, y) = u(x|y)$, so $\nabla u(x_{n+1}) - \nabla u(x_n) = -\nabla f(x_n)$.

iii) Natural gradient descent: $c(x, y) = u(y|x)$, so $x_{n+1} - x_n = -(\nabla^2 u(x_n))^{-1}\nabla f(x_n)$.

iv) A nonlinear gradient descent: $c(x, y) = \ell(x - y)$, so $x_{n+1} - x_n = -\nabla \ell^*(\nabla f(x_n))$.

v) Riemannian gradient descent: (M, g) a Riemannian manifold. Take $X = Y = M$ and $c(x, y) = \frac{L}{2}d^2(x, y)$, so $x_{n+1} = \exp_{x_n}(-\frac{1}{L}\nabla f(x_n))$,

Gradient descent with a general cost - Examples

$$-\nabla_x c(x_n, y_{n+1}) = -\nabla f(x_n),$$
$$\nabla_x c(x_{n+1}, y_{n+1}) = 0.$$

In the following: $Y = X$, and c is minimal on the diagonal $\{x = y\}$, so $x_{n+1} = y_{n+1}$

i) Gradient descent: $c(x, y) = \frac{L}{2}\|x - y\|^2$ and $x_{n+1} - x_n = -\frac{1}{L}\nabla f(x_n)$.

ii) Mirror descent: $c(x, y) = u(x|y)$, so $\nabla u(x_{n+1}) - \nabla u(x_n) = -\nabla f(x_n)$.

iii) Natural gradient descent: $c(x, y) = u(y|x)$, so $x_{n+1} - x_n = -(\nabla^2 u(x_n))^{-1}\nabla f(x_n)$.

iv) A nonlinear gradient descent: $c(x, y) = \ell(x - y)$, so $x_{n+1} - x_n = -\nabla \ell^*(\nabla f(x_n))$.

v) Riemannian gradient descent: (M, g) a Riemannian manifold. Take $X = Y = M$ and $c(x, y) = \frac{L}{2}d^2(x, y)$, so $x_{n+1} = \exp_{x_n}(-\frac{1}{L}\nabla f(x_n))$,

Cool, but what do you need to converge?

$\hookrightarrow$ Something like L-smoothness and μ-strong convexity

c-cross-convexity

Consider the sequence of AM iterates, starting from any x_0,

$$y_n \to x_n \to y_{n+1}$$

We say that f is λ-strongly c-cross-convex for $\lambda \geq 0$ if, for all $x, y_n \in X \times Y$,

$$f(x) - f(x_n) \geq c(x, y_{n+1}) - c(x, y_n) + c(x_n, y_n) - c(x_n, y_{n+1}) + \lambda(c(x, y_n) - c(x_n, y_n))..$$

c-concavity ($f(x) = \inf_y c(x, y) + f^c(y)$) implies, since $f^c(y_{n+1}) = f(x_n) - c(x_n, y_{n+1})$,

$$f(x) - f(x_n) \leq c(x, y_{n+1}) - c(x_n, y_{n+1}).$$

These conditions extend L-smoothness and (strong) convexity when $c(x, y) = \frac{L}{2}\|x - y\|^2$.[5]

[5] Flavien Léger and Pierre-Cyril Aubin-Frankowski. "Gradient descent with a general cost". In: (2023). https://arxiv.org/abs/2305.04917.

Theorem (Convergence rates for gradient descent with general cost)

i) Suppose that f is c-concave. Then we have the descent property+stopping criterion

$$f(x_{n+1}) \leq f(x_n) - [c(x_n, y_{n+1}) - c(x_{n+1}, y_{n+1})] \leq f(x_n),$$

$$\min_{0\leq k\leq n-1} [c(x_k, y_{k+1}) - c(x_{k+1}, y_{k+1})] \leq \frac{f(x_0) - f_*}{n}.$$

ii) Suppose in addition that f is c-cross-convex. Then for any $x \in X, n \geq 1$,

$$f(x_n) \leq f(x) + \frac{c(x, y_0) - c(x_0, y_0)}{n}. \tag{17}$$

iii) Suppose in addition that f is λ-strongly c-cross-convex for some $\lambda \in (0, 1)$. Then for any $x \in X, n \geq 1$, setting $\Lambda := (1 - \lambda)^{-1} > 1$

$$f(x_n) \leq f(x) + \frac{\lambda (c(x, y_0) - c(x_0, y_0))}{\Lambda^n - 1}. \tag{18}$$

What have we seen? What can you see more in the articles?

Linear optimal control/estimation duality

LQ optimal control $\subset$ kernel methods. New formulas for the covariances of GPs induced by linear SDEs!

Global optimization of smooth functions

Kernel Sum-of-Squares use smoothness against curse of dimensionality!

Tropical kernels

Representer theorems still hold in max-plus settings! There are also analogies with Hilbertian framework and applications to value functions.

c-concavity for revisiting optimization algorithms!

c-concavity and c-cross-convexity generalize smoothness and convexity and encompass many algorithms! New assumptions for global convergence of natural gradient descent/Newton

What have we seen? What can you see more in the articles?

Linear optimal control/estimation duality

LQ optimal control $\subset$ kernel methods. New formulas for the covariances of GPs induced by linear SDEs!

Global optimization of smooth functions

Kernel Sum-of-Squares use smoothness against curse of dimensionality!

Thank you for your attention!

Tropical kernels

Representer theorems still hold in max-plus settings! There are also analogies with Hilbertian framework and applications to value functions.

***c*-concavity for revisiting optimization algorithms!**

c-concavity and c-cross-convexity generalize smoothness and convexity and encompass many algorithms! New assumptions for global convergence of natural gradient descent/Newton

References I

- Aubin-Frankowski, Pierre-Cyril. "Linearly Constrained Linear Quadratic Regulator from the Viewpoint of Kernel Methods". In: *SIAM Journal on Control and Optimization* 59.4 (2021), pp. 2693–2716.
- Aubin-Frankowski, Pierre-Cyril and Alain Bensoussan. "The reproducing kernel Hilbert spaces underlying linear SDE Estimation, Kalman filtering and their relation to optimal control". In: *Pure and Applied Functional Analysis* (2022).
- Aubin-Frankowski, Pierre-Cyril and Stéphane Gaubert. "Tropical reproducing kernels and optimization". In: *Integral Equations and Operator Theory* (2023). (to be published).
- Aubin-Frankowski, Pierre-Cyril and Alessandro Rudi. "Approximation of optimization problems with constraints through kernel Sum-Of-Squares". In: (2022). https://arxiv.org/abs/2301.06339.
- Boyd, Stephen and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. DOI: 10.1017/CBO9780511804441.

References II

- Léger, Flavien and Pierre-Cyril Aubin-Frankowski. "Gradient descent with a general cost". In: (2023). https://arxiv.org/abs/2305.04917.
- Rudi, Alessandro, Ulysse Marteau-Ferey, and Francis Bach. *Finding Global Minima via Kernel Approximations*. 2020. arXiv: 2012.11978 [math.OC].
- Schölkopf, B., R. Herbrich, and A. J. Smola. "A Generalized Representer Theorem". In: *Computational Learning Theory (CoLT)*. 2001, pp. 416–426.