Kernel Machines with Hard Shape Constraints

Zoltán Szabó

Joint work with: Pierre-Cyril Aubin-Frankowski @ MINES ParisTech

Meeting on Mathematical Statistics (MMS),
Robustness and Computational Efficiency of Algorithms in Statistical Learning
December 15, 2020

Shape constraints

Pattern

$$0 \leq Df(x) \quad \forall x.$$

Shape constraints

Pattern

$$0 \leq Df(x) \quad \forall x.$$

Examples:

1. non-negativity: $\boxed{0 \leq f(x)}$,

Shape constraints

Pattern

$$0 \leq Df(\mathrm{x}) \quad \forall \mathrm{x}.$$

Examples:

1. non-negativity: $\boxed{0 \leq f(x)}$,
2. monotonicity ($\nearrow$): $\boxed{0 \leq f'(x)}$,

Shape constraints

Pattern

$$0 \leq Df(\mathrm{x}) \quad \forall \mathrm{x}.$$

Examples:

1. non-negativity: $\boxed{0 \leq f(x)}$,
2. monotonicity ($\nearrow$): $\boxed{0 \leq f'(x)}$,
3. convexity: $\boxed{0 \leq f''(x)}$,

Shape constraints

Pattern

$$0 \le Df(x) \quad \forall x.$$

Examples:

1. non-negativity: $\boxed{0 \le f(x)}$,
2. monotonicity ($\nearrow$): $\boxed{0 \le f'(x)}$,
3. convexity: $\boxed{0 \le f''(x)}$,
4. n-monotonicity: $\boxed{0 \le f^{(n)}(x)}$,

Shape constraints

Pattern

$$0 \leq Df(x) \quad \forall x.$$

Examples:

1. non-negativity: $\boxed{0 \leq f(x)}$,
2. monotonicity ($\nearrow$): $\boxed{0 \leq f'(x)}$,
3. convexity: $\boxed{0 \leq f''(x)}$,
4. n-monotonicity: $\boxed{0 \leq f^{(n)}(x)}$,
5. $(n-1)$-alternating monotonicity: for $n \geq 2$

$$\boxed{(-1)^j f^{(j)}} : \boxed{\geq 0}, \boxed{\nearrow} \text{ and } \boxed{\text{convex}} \ \forall j \in [\![0, n-2]\!].$$

Example: generator of a d-variate Archimedean copula is $(d-2)$-alternating monotone.

Examples continued

6. Monotonicity w.r.t. partial ordering ($\mathrm{u} \preccurlyeq \mathrm{v} \Rightarrow f(\mathrm{u}) \leq f(\mathrm{v})$):

$\mathrm{u} \preccurlyeq \mathrm{v}$ iff

- $u_i \leq v_i$ ($\forall i$; product ordering),
- $\sum_{j \in [i]} u_j \leq \sum_{j \in [i]} v_j$ ($\forall i$; unordered weak majorization).

Examples continued

6. Monotonicity w.r.t. partial ordering ($\mathbf{u} \preccurlyeq \mathbf{v} \Rightarrow f(\mathbf{u}) \leq f(\mathbf{v})$):

$$\boxed{0 \leq \partial^{\mathbf{e}_j} f(\mathbf{x})}, \quad (\forall j \in [d], \forall \mathbf{x}),$$

$$\boxed{0 \leq \partial^{\mathbf{e}_d} f(\mathbf{x}) \leq \ldots \leq \partial^{\mathbf{e}_1} f(\mathbf{x})} \quad (\forall \mathbf{x}).$$

$\mathbf{u} \preccurlyeq \mathbf{v}$ iff

- $u_i \leq v_i$ ($\forall i$; product ordering),
- $\sum_{j \in [i]} u_j \leq \sum_{j \in [i]} v_j$ ($\forall i$; unordered weak majorization).

Examples continued

6. Monotonicity w.r.t. partial ordering ($\mathrm{u} \preccurlyeq \mathrm{v} \Rightarrow f(\mathrm{u}) \leq f(\mathrm{v})$):

$$\boxed{0 \leq \partial^{\mathrm{e}_j} f(\mathrm{x})}, \quad (\forall j \in [d], \forall \mathrm{x}),$$

$$\boxed{0 \leq \partial^{\mathrm{e}_d} f(\mathrm{x}) \leq \ldots \leq \partial^{\mathrm{e}_1} f(\mathrm{x})} \quad (\forall \mathrm{x}).$$

$\mathrm{u} \preccurlyeq \mathrm{v}$ iff

- $u_i \leq v_i$ ($\forall i$; product ordering),
- $\sum_{j \in [i]} u_j \leq \sum_{j \in [i]} v_j$ ($\forall i$; unordered weak majorization).

7. Supermodularity:

$$\boxed{0 \leq \frac{\partial^2 f(\mathrm{x})}{\partial x_i \partial x_j}} \quad (\forall i \neq j \in [d], \forall \mathrm{x}),$$

i.e. $f(\mathrm{u} \vee \mathrm{v}) + f(\mathrm{u} \wedge \mathrm{v}) \geq f(\mathrm{u}) + f(\mathrm{v})$ for all $\mathrm{u}, \mathrm{v} \in \mathbb{R}^d$.

Shape constraints are omnipresent [Johnson and Jiang, 2018, Guntuboyina and Sen, 2018, Chetverikov et al., 2018]

- Economics:
  - utility functions are $\nearrow$ and concave [Matzkin, 1991].

Shape constraints are omnipresent [Johnson and Jiang, 2018, Guntuboyina and Sen, 2018, Chetverikov et al., 2018]

- Economics:
  - utility functions are $\boxed{\nearrow}$ and $\boxed{\text{concave}}$ [Matzkin, 1991].
  - demand functions of normal goods are $\boxed{\text{downward sloping}}$ [Lewbel, 2010, Blundell et al., 2012],

Shape constraints are omnipresent [Johnson and Jiang, 2018, Guntuboyina and Sen, 2018, Chetverikov et al., 2018]

- Economics:
  - utility functions are $\nearrow$ and concave [Matzkin, 1991].
  - demand functions of normal goods are downward sloping [Lewbel, 2010, Blundell et al., 2012],
  - production functions are concave [Varian, 1984] or S-shaped [Yagi et al., 2020].

Shape constraints are omnipresent [Johnson and Jiang, 2018, Guntuboyina and Sen, 2018, Chetverikov et al., 2018]

- Economics:
  - utility functions are $\nearrow$ and concave [Matzkin, 1991].
  - demand functions of normal goods are downward sloping [Lewbel, 2010, Blundell et al., 2012],
  - production functions are concave [Varian, 1984] or S-shaped [Yagi et al., 2020].
  - panel multinomial choice problems [Shi et al., 2018]: cyclic monotonicity,

Shape constraints are omnipresent [Johnson and Jiang, 2018, Guntuboyina and Sen, 2018, Chetverikov et al., 2018]

- Economics:
  - utility functions are $\nearrow$ and concave [Matzkin, 1991].
  - demand functions of normal goods are downward sloping [Lewbel, 2010, Blundell et al., 2012],
  - production functions are concave [Varian, 1984] or S-shaped [Yagi et al., 2020].
  - panel multinomial choice problems [Shi et al., 2018]: cyclic monotonicity,
  - single index model: most link functions are monotone [Li and Racine, 2007, Chen and Samworth, 2016, Balabdaoui et al., 2019].

Shape constraints are omnipresent [Johnson and Jiang, 2018, Guntuboyina and Sen, 2018, Chetverikov et al., 2018]

- Economics:
  - utility functions are $\nearrow$ and concave [Matzkin, 1991].
  - demand functions of normal goods are downward sloping [Lewbel, 2010, Blundell et al., 2012],
  - production functions are concave [Varian, 1984] or S-shaped [Yagi et al., 2020].
  - panel multinomial choice problems [Shi et al., 2018]: cyclic monotonicity,
  - single index model: most link functions are monotone [Li and Racine, 2007, Chen and Samworth, 2016, Balabdaoui et al., 2019].
- Biology (monotone regression): identify genome interactions [Luss et al., 2012], dose-response studies [Hu et al., 2005].

Few applications

- Statistics: quantile function $\nearrow$ w.r.t. the quantile level, pdfs are non-negative and often log-concave.

Few applications

- Statistics: quantile function $\boxed{\nearrow}$ w.r.t. the quantile level, pdfs are non-negative and often log-concave.
- Finance:
  - European and American call option prices: convex & monotone in the underlying stock price and $\boxed{\nearrow}$ in volatility [Aït-Sahalia and Duarte, 2003].

Few applications

- Statistics: quantile function $\nearrow$ w.r.t. the quantile level, pdfs are non-negative and often log-concave.
- Finance:
  - European and American call option prices: convex & monotone in the underlying stock price and $\nearrow$ in volatility [Aït-Sahalia and Duarte, 2003].
- RL and stochastic optimization: value functions are often convex [Keshavarz et al., 2011, Shapiro et al., 2014].

Few applications

- Statistics: quantile function $\boxed{\nearrow}$ w.r.t. the quantile level, pdfs are non-negative and often log-concave.
- Finance:
  - European and American call option prices: convex & monotone in the underlying stock price and $\boxed{\nearrow}$ in volatility [Aït-Sahalia and Duarte, 2003].
- RL and stochastic optimization: value functions are often convex [Keshavarz et al., 2011, Shapiro et al., 2014].
- Supply chain models, stochastic multi-period inventory problems, pricing models and game theory: supermodularity [Topkis, 1998, Simchi-Levi et al., 2014].

Typical setting: supervised learning

- Find $f \in \mathcal{H}$ such that $f(\mathrm{x}_n) \approx y_n$, $0 \leq Df(\mathrm{x})$ $\forall \mathrm{x} \in K$.

Typical setting: supervised learning

- Find $f \in \mathcal{H}$ such that $f(\mathrm{x}_n) \approx y_n$, $0 \leq Df(\mathrm{x})$ $\forall \mathrm{x} \in K$.
- Various exciting approaches with asymptotic guarantees, but
  1. they are often 'soft': restriction at finite many points,

Typical setting: supervised learning

- Find $f \in \mathcal{H}$ such that $f(\mathrm{x}_n) \approx y_n$, $0 \leq Df(\mathrm{x})$ $\forall \mathrm{x} \in K$.
- Various exciting approaches with asymptotic guarantees, but
  1. they are often 'soft': restriction at finite many points,
  2. use simplistic function classes: polynomials, polynomial splines,

Typical setting: supervised learning

- Find $f \in \mathcal{H}$ such that $f(\mathrm{x}_n) \approx y_n$, $0 \leq Df(\mathrm{x})$ $\forall \mathrm{x} \in K$.
- Various exciting approaches with asymptotic guarantees, <u>but</u>
  1. they are often 'soft': restriction at finite many points,
  2. use simplistic function classes: polynomials, polynomial splines,
  3. apply hard-wired parameterizations: exponential, quadratic, or

Typical setting: supervised learning

- Find $f \in \mathcal{H}$ such that $f(\mathrm{x}_n) \approx y_n$, $0 \le Df(\mathrm{x})$ $\forall \mathrm{x} \in K$.
- Various exciting approaches with asymptotic guarantees, but
  1. they are often 'soft': restriction at finite many points,
  2. use simplistic function classes: polynomials, polynomial splines,
  3. apply hard-wired parameterizations: exponential, quadratic, or
  4. only work for (a few) fixed Ds.

Typical setting: supervised learning

- Find $f \in \mathcal{H}$ such that $f(\mathrm{x}_n) \approx y_n$, $0 \leq Df(\mathrm{x})$ $\forall \mathrm{x} \in K$.
- Various exciting approaches with asymptotic guarantees, but
  1. they are often 'soft': restriction at finite many points,
  2. use simplistic function classes: polynomials, polynomial splines,
  3. apply hard-wired parameterizations: exponential, quadratic, or
  4. only work for (a few) fixed Ds.

Today: optimization framework

rich $\mathcal{H}$, hard ($\forall \mathrm{x} \in K$) shape constraints, modularity in D.

Typical setting: supervised learning

- Find $f \in \mathcal{H}$ such that $f(\mathrm{x}_n) \approx y_n$, $0 \leq Df(\mathrm{x})$ $\forall \mathrm{x} \in K$.
- Various exciting approaches with asymptotic guarantees, but
  1. they are often 'soft': restriction at finite many points,
  2. use simplistic function classes: polynomials, polynomial splines,
  3. apply hard-wired parameterizations: exponential, quadratic, or
  4. only work for (a few) fixed Ds.

Today: optimization framework

rich $\mathcal{H}$, hard ($\forall \mathrm{x} \in K$) shape constraints, modularity in D.

Towards flexible $\mathcal{H}$-s . . .

Kernel motivation

- In $\mathbb{R}^d$: $\langle \mathrm{x}, \mathrm{x}' \rangle = \sum_{i \in [d]} x_i x_i' \Rightarrow \left\| \mathrm{x} - \mathrm{x}' \right\|_2$ and $\sphericalangle\left(\mathrm{x}, \mathrm{x}'\right)$.

Kernel motivation

- In $\mathbb{R}^d$: $\langle \mathrm{x}, \mathrm{x}' \rangle = \sum_{i \in [d]} x_i x_i' \Rightarrow \left\| \mathrm{x} - \mathrm{x}' \right\|_2$ and $\sphericalangle\left(\mathrm{x}, \mathrm{x}'\right)$.
- Nonlinear features ($d = 2$):

$$\varphi(\mathrm{x}) = \left(x_1^2, \sqrt{2} x_1 x_2, x_2^2\right),$$

$$\boxed{\langle \varphi(\mathrm{x}), \varphi(\mathrm{x}') \rangle} = \left\langle \begin{bmatrix} x_1^2 \\ \sqrt{2} x_1 x_2 \\ x_2^2 \end{bmatrix}, \begin{bmatrix} (x_1')^2 \\ \sqrt{2} (x_1')(x_2') \\ (x_2')^2 \end{bmatrix} \right\rangle$$

Kernel motivation

- In $\mathbb{R}^d$: $\langle \mathrm{x}, \mathrm{x}' \rangle = \sum_{i \in [d]} x_i x_i' \Rightarrow \left\| \mathrm{x} - \mathrm{x}' \right\|_2$ and $\sphericalangle \left(\mathrm{x}, \mathrm{x}' \right)$.
- Nonlinear features ($d = 2$):

$$\varphi(\mathrm{x}) = \left(x_1^2, \sqrt{2} x_1 x_2, x_2^2 \right),$$

$$\boxed{\langle \varphi(\mathrm{x}), \varphi(\mathrm{x}') \rangle} = \left\langle \begin{bmatrix} x_1^2 \\ \sqrt{2} x_1 x_2 \\ x_2^2 \end{bmatrix}, \begin{bmatrix} (x_1')^2 \\ \sqrt{2} (x_1')(x_2') \\ (x_2')^2 \end{bmatrix} \right\rangle$$

$$= x_1^2 (x_1')^2 + \underbrace{\sqrt{2}\sqrt{2}}_{2} x_1 x_2 (x_1')(x_2') + x_2^2 (x_2')^2$$

Kernel motivation

- In $\mathbb{R}^d$: $\langle \mathrm{x}, \mathrm{x}' \rangle = \sum_{i \in [d]} x_i x_i' \Rightarrow \left\| \mathrm{x} - \mathrm{x}' \right\|_2$ and $\sphericalangle(\mathrm{x}, \mathrm{x}')$.
- Nonlinear features ($d = 2$):

$$\begin{aligned}
\varphi(\mathrm{x}) &= \left(x_1^2, \sqrt{2} x_1 x_2, x_2^2 \right), \\
\boxed{\langle \varphi(\mathrm{x}), \varphi(\mathrm{x}') \rangle} &= \left\langle \begin{bmatrix} x_1^2 \\ \sqrt{2} x_1 x_2 \\ x_2^2 \end{bmatrix}, \begin{bmatrix} (x_1')^2 \\ \sqrt{2}(x_1')(x_2') \\ (x_2')^2 \end{bmatrix} \right\rangle \\
&= x_1^2 (x_1')^2 + \underbrace{\sqrt{2}\sqrt{2}}_{2} x_1 x_2 (x_1')(x_2') + x_2^2 (x_2')^2 \\
&= \left(x_1 x_1' + x_2 x_2' \right)^2
\end{aligned}$$

Kernel motivation

- In $\mathbb{R}^d$: $\langle \mathbf{x}, \mathbf{x}' \rangle = \sum_{i\in[d]} x_i x_i' \Rightarrow \left\| \mathbf{x} - \mathbf{x}' \right\|_2$ and $\sphericalangle(\mathbf{x}, \mathbf{x}')$.
- Nonlinear features ($d = 2$):

$$\varphi(\mathbf{x}) = \left(x_1^2, \sqrt{2}x_1x_2, x_2^2\right),$$

$$\boxed{\langle \varphi(\mathbf{x}), \varphi(\mathbf{x}') \rangle} = \left\langle \begin{bmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \end{bmatrix}, \begin{bmatrix} (x_1')^2 \\ \sqrt{2}(x_1')(x_2') \\ (x_2')^2 \end{bmatrix} \right\rangle$$

$$= x_1^2(x_1')^2 + \underbrace{\sqrt{2}\sqrt{2}}_{2} x_1x_2(x_1')(x_2') + x_2^2(x_2')^2$$

$$= \left(x_1x_1' + x_2x_2'\right)^2$$

$$= \left\langle \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \begin{bmatrix} x_1' \\ x_2' \end{bmatrix} \right\rangle^2 = \boxed{\langle \mathbf{x}, \mathbf{x}' \rangle^2} =: k(\mathbf{x}, \mathbf{x}').$$

Kernel motivation

- In $\mathbb{R}^d$: $\langle \mathrm{x}, \mathrm{x}' \rangle = \sum_{i\in[d]} x_i x_i' \Rightarrow \left\| \mathrm{x} - \mathrm{x}' \right\|_2$ and $\sphericalangle(\mathrm{x}, \mathrm{x}')$.
- Nonlinear features ($d = 2$):

$$
\begin{aligned}
\varphi(\mathrm{x}) &= \left(x_1^2, \sqrt{2} x_1 x_2, x_2^2\right), \\
\boxed{\langle \varphi(\mathrm{x}), \varphi(\mathrm{x}') \rangle} &= \left\langle \begin{bmatrix} x_1^2 \\ \sqrt{2} x_1 x_2 \\ x_2^2 \end{bmatrix}, \begin{bmatrix} (x_1')^2 \\ \sqrt{2}(x_1')(x_2') \\ (x_2')^2 \end{bmatrix} \right\rangle \\
&= x_1^2 (x_1')^2 + \underbrace{\sqrt{2}\sqrt{2}}_{2} x_1 x_2 (x_1')(x_2') + x_2^2 (x_2')^2 \\
&= \left(x_1 x_1' + x_2 x_2'\right)^2 \\
&= \left\langle \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \begin{bmatrix} x_1' \\ x_2' \end{bmatrix} \right\rangle^2 = \boxed{\langle \mathrm{x}, \mathrm{x}' \rangle^2} =: k(\mathrm{x}, \mathrm{x}').
\end{aligned}
$$

$\langle \mathrm{x}, \mathrm{x}' \rangle^d = \langle \varphi(\mathrm{x}), \varphi(\mathrm{x}') \rangle$: $\varphi(\mathrm{x}) =$ d-order polynomial. $\Rightarrow$

Kernel motivation

- In $\mathbb{R}^d$: $\langle \mathbf{x}, \mathbf{x}' \rangle = \sum_{i \in [d]} x_i x_i' \Rightarrow \left\| \mathbf{x} - \mathbf{x}' \right\|_2$ and $\sphericalangle(\mathbf{x}, \mathbf{x}')$.
- Nonlinear features ($d = 2$):

$$
\begin{aligned}
\varphi(\mathbf{x}) &= \left(x_1^2, \sqrt{2} x_1 x_2, x_2^2 \right), \\
\boxed{\langle \varphi(\mathbf{x}), \varphi(\mathbf{x}') \rangle} &= \left\langle \begin{bmatrix} x_1^2 \\ \sqrt{2} x_1 x_2 \\ x_2^2 \end{bmatrix}, \begin{bmatrix} (x_1')^2 \\ \sqrt{2}(x_1')(x_2') \\ (x_2')^2 \end{bmatrix} \right\rangle \\
&= x_1^2 (x_1')^2 + \underbrace{\sqrt{2}\sqrt{2}}_{2} x_1 x_2 (x_1')(x_2') + x_2^2 (x_2')^2 \\
&= \left(x_1 x_1' + x_2 x_2' \right)^2 \\
&= \left\langle \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \begin{bmatrix} x_1' \\ x_2' \end{bmatrix} \right\rangle^2 = \boxed{\langle \mathbf{x}, \mathbf{x}' \rangle^2} =: k(\mathbf{x}, \mathbf{x}').
\end{aligned}
$$

$\langle \mathbf{x}, \mathbf{x}' \rangle^d = \langle \varphi(\mathbf{x}), \varphi(\mathbf{x}') \rangle$: $\varphi(\mathbf{x}) = d$-order polynomial. $\Rightarrow$ Explicit computation would be heavy!

Kernel, RKHS: definition

- Def-1 (feature space):

$$k(x, y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}} .$$

Kernel, RKHS: definition

- Def-1 (feature space):

$$k(x, y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}}.$$

- Def-2 (evaluation): $\delta_x(f) = f(x)$ is continuous for all x.

Kernel, RKHS: definition

- Def-1 (feature space):

$$k(x, y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}}.$$

- Def-2 (evaluation): $\delta_x(f) = f(x)$ is continuous for all x.
- Def-3 (Gram matrix): $\mathrm{G} = [k(x_i, x_j)]_{i,j=1}^n \succeq 0$.

Kernel, RKHS: definition

- Def-1 (feature space):

$$k(x,y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}}.$$

- Def-2 (evaluation): $\delta_x(f) = f(x)$ is continuous for all x.
- Def-3 (Gram matrix): $\mathrm{G} = [k(x_i, x_j)]_{i,j=1}^n \succeq 0$.
- Def-4 (reproducing kernel):

$$k(\cdot, x) \in \mathcal{H}, \qquad f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}}.$$

Constructively, $\mathcal{H}_k = \overline{\left\{\sum_{i=1}^n \alpha_i k(\cdot, x_i) \,:\, n \in \mathbb{N}, x_i \in \mathcal{X}\right\}}$.

Kernel, RKHS: definition

- Def-1 (feature space):

$$k(x, y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}}.$$

- Def-2 (evaluation): $\delta_x(f) = f(x)$ is continuous for all x.
- Def-3 (Gram matrix): $\mathbf{G} = [k(x_i, x_j)]_{i,j=1}^n \succeq 0$.
- Def-4 (reproducing kernel):

$$k(\cdot, x) \in \mathcal{H}, \qquad f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}}.$$

Constructively, $\mathcal{H}_k = \overline{\left\{\sum_{i=1}^n \alpha_i k(\cdot, x_i) \,:\, n \in \mathbb{N}, x_i \in \mathcal{X}\right\}}$.

- All these definitions are equivalent, $k \overset{1:1}{\leftrightarrow} \mathcal{H}_k$.
- Included: Fourier analysis, polynomials, splines, ...

Kernel examples on $\mathbb{R}^d$ ($\gamma, \sigma, v > 0$, $c \geq 0$, $p \in \mathbb{Z}^+$)

$$k_p(\mathrm{x}, \mathrm{y}) = (\langle \mathrm{x}, \mathrm{y} \rangle + c)^p$$

Kernel examples on $\mathbb{R}^d$ ($\gamma, \sigma, v > 0$, $c \geq 0$, $p \in \mathbb{Z}^+$)

$$k_p(\mathrm{x}, \mathrm{y}) = (\langle \mathrm{x}, \mathrm{y} \rangle + c)^p, \qquad k_G(\mathrm{x}, \mathrm{y}) = e^{-\gamma \|\mathrm{x}-\mathrm{y}\|_2^2},$$

$$k_e(\mathrm{x}, \mathrm{y}) = e^{-\gamma \|\mathrm{x}-\mathrm{y}\|_2}, \qquad k_L(\mathrm{x}, \mathrm{y}) = e^{-\gamma \|\mathrm{x}-\mathrm{y}\|_1},$$

$$k_C(\mathrm{x}, \mathrm{y}) = \frac{1}{1 + \gamma \|\mathrm{x} - \mathrm{y}\|_2^2}, \qquad k_{\tilde{e}}(\mathrm{x}, \mathrm{y}) = e^{\gamma \langle \mathrm{x}, \mathrm{y} \rangle}.$$

Kernel examples on $\mathbb{R}^d$ ($\gamma, \sigma, v > 0$, $c \geq 0$, $p \in \mathbb{Z}^+$)

$$k_p(\mathrm{x},\mathrm{y}) = (\langle \mathrm{x},\mathrm{y}\rangle + c)^p, \qquad k_G(\mathrm{x},\mathrm{y}) = e^{-\gamma\|\mathrm{x}-\mathrm{y}\|_2^2},$$

$$k_e(\mathrm{x},\mathrm{y}) = e^{-\gamma\|\mathrm{x}-\mathrm{y}\|_2}, \qquad k_L(\mathrm{x},\mathrm{y}) = e^{-\gamma\|\mathrm{x}-\mathrm{y}\|_1},$$

$$k_C(\mathrm{x},\mathrm{y}) = \frac{1}{1+\gamma\|\mathrm{x}-\mathrm{y}\|_2^2}, \qquad k_{\tilde{e}}(\mathrm{x},\mathrm{y}) = e^{\gamma\langle \mathrm{x},\mathrm{y}\rangle}.$$

Or the flexible Matérn family:

$$k_M(\mathrm{x},\mathrm{y}) = \frac{2^{1-v}}{\Gamma(v)}\left(\frac{\sqrt{2v}\,\|\mathrm{x}-\mathrm{y}\|_2}{\sigma}\right)^v K_v\left(\frac{\sqrt{2v}\,\|\mathrm{x}-\mathrm{y}\|_2}{\sigma}\right),$$

where

- K_v: modified Bessel function of the second kind of order v,
- Specific cases: For $v = \frac{1}{2}$ one gets $k(\mathrm{x},\mathrm{y}) = e^{-\frac{\|\mathrm{x}-\mathrm{y}\|_2}{\sigma}}$.
  Gaussian kernel: $v \to \infty$.

Kernels on other domains ($\mathcal{X}$)

- Strings [Watkins, 1999, Lodhi et al., 2002, Leslie et al., 2002, Kuang et al., 2004, Leslie and Kuang, 2004, Saigo et al., 2004, Cuturi and Vert, 2005],
- time series [Rüping, 2001, Cuturi et al., 2007, Cuturi, 2011, Király and Oberhauser, 2019],
- trees [Collins and Duffy, 2001, Kashima and Koyanagi, 2002],
- groups and specifically rankings [Cuturi et al., 2005, Jiao and Vert, 2016],
- sets [Haussler, 1999, Gärtner et al., 2002],
- various generative models [Jaakkola and Haussler, 1999, Tsuda et al., 2002, Seeger, 2002, Jebara et al., 2004],
- fuzzy domains [Guevara et al., 2017], or
- graphs [Kondor and Lafferty, 2002, Gärtner et al., 2003, Kashima et al., 2003, Borgwardt and Kriegel, 2005, Shervashidze et al., 2009, Vishwanathan et al., 2010, Kondor and Pan, 2016, Bai et al., 2020].

Why kernel and RKHSs?

1. Numerous data types.

Why kernel and RKHSs?

1. Numerous data types.
2. RKHS can
   - be dense in various function spaces [Steinwart, 2001, Micchelli et al., 2006, Sriperumbudur et al., 2011, Simon-Gabriel and Schölkopf, 2018],
   - encode probability measures injectively [Fukumizu et al., 2008, Sriperumbudur et al., 2010]

$$\mathbb{P} \mapsto \int_{\mathcal{X}} \varphi(x)\mathrm{d}\mathbb{P}(x) \in \mathcal{H}_k$$

   - characterize independence of random variables [Bach and Jordan, 2002, Blanchard et al., 2011, Gretton, 2015, Szabó and Sriperumbudur, 2018].

Why kernel and RKHSs?

1. Numerous data types.
2. RKHS can
   - be dense in various function spaces [Steinwart, 2001, Micchelli et al., 2006, Sriperumbudur et al., 2011, Simon-Gabriel and Schölkopf, 2018],
   - encode probability measures injectively [Fukumizu et al., 2008, Sriperumbudur et al., 2010]

$$\mathbb{P} \mapsto \int_{\mathcal{X}} \varphi(x) \mathrm{d}\mathbb{P}(x) \in \mathcal{H}_k$$

   - characterize independence of random variables [Bach and Jordan, 2002, Blanchard et al., 2011, Gretton, 2015, Szabó and Sriperumbudur, 2018].
3. Computationally tractable: $k(x_i, x_j)$.

Why kernel and RKHSs?

1. Numerous data types.
2. RKHS can
   - be dense in various function spaces [Steinwart, 2001, Micchelli et al., 2006, Sriperumbudur et al., 2011, Simon-Gabriel and Schölkopf, 2018],
   - encode probability measures injectively [Fukumizu et al., 2008, Sriperumbudur et al., 2010]

$$\mathbb{P} \mapsto \int_{\mathcal{X}} \varphi(x) \mathrm{d}\mathbb{P}(x) \in \mathcal{H}_k$$

   - characterize independence of random variables [Bach and Jordan, 2002, Blanchard et al., 2011, Gretton, 2015, Szabó and Sriperumbudur, 2018].
3. Computationally tractable: $k(x_i, x_j)$.
4. Hilbert structure $\Rightarrow$ statistical analysis.

Why kernel and RKHSs?

1. Numerous data types.
2. RKHS can
   - be dense in various function spaces [Steinwart, 2001, Micchelli et al., 2006, Sriperumbudur et al., 2011, Simon-Gabriel and Schölkopf, 2018],
   - encode probability measures injectively [Fukumizu et al., 2008, Sriperumbudur et al., 2010]

$$\mathbb{P} \mapsto \int_{\mathcal{X}} \varphi(x) \mathrm{d}\mathbb{P}(x) \in \mathcal{H}_k$$

   - characterize independence of random variables [Bach and Jordan, 2002, Blanchard et al., 2011, Gretton, 2015, Szabó and Sriperumbudur, 2018].
3. Computationally tractable: $k(x_i, x_j)$.
4. Hilbert structure $\Rightarrow$ statistical analysis.
5. Vector-valued RKHSs [Pedrick, 1957, Micchelli and Pontil, 2005, Carmeli et al., 2006].

Task-1: joint quantile regression (JQR)

- Given: $(\tau_q)_{q\in[Q]} \subset (0,1)$ levels $\nearrow$, $\{(\mathrm{x}_n, y_n)\}_{n\in[N]}$ samples.
- Estimate jointly the τ_q-quantiles of $\mathbb{P}(Y|X=\mathrm{x})$.

Task-1: joint quantile regression (JQR)

- Given: $(\tau_q)_{q\in[Q]} \subset (0,1)$ levels $\nearrow$, $\{(\mathrm{x}_n, y_n)\}_{n\in[N]}$ samples.
- Estimate jointly the τ_q-quantiles of $\mathbb{P}(Y|X=\mathrm{x})$[Sangnier et al., 2016].
- Objective:

$$\mathcal{L}(\mathrm{f},\mathrm{b}) = \frac{1}{N}\sum_{q\in[Q]}\sum_{n\in[N]} l_{\tau_q}\left(y_n - [f_q(\mathrm{x}_n) + b_q]\right) + \lambda_{\mathrm{b}}\|\mathrm{b}\|_2^2 + \lambda_f \sum_{q\in[Q]} \|f_q\|_k^2,$$

$$l_\tau(e) = \max(\tau e, (\tau-1)e).$$

Task-1: joint quantile regression (JQR)

- Given: $(\tau_q)_{q\in[Q]} \subset (0,1)$ levels $\nearrow$, $\{(\mathrm{x}_n, y_n)\}_{n\in[N]}$ samples.
- Estimate jointly the τ_q-quantiles of $\mathbb{P}(Y|X=\mathrm{x})$[Sangnier et al., 2016].
- Objective:

$$\mathcal{L}(\mathrm{f},\mathrm{b}) = \frac{1}{N}\sum_{q\in[Q]}\sum_{n\in[N]} l_{\tau_q}\left(y_n - [f_q(\mathrm{x}_n) + b_q]\right) + \lambda_{\mathrm{b}}\|\mathrm{b}\|_2^2 + \lambda_f \sum_{q\in[Q]} \|f_q\|_k^2,$$

$$l_\tau(e) = \max(\tau e, (\tau-1)e).$$

- Constraint (non-crossing): $K :=$ smallest rectangle containing $\{\mathrm{x}_n\}_{n\in[N]}$,

$$f_q(\mathrm{x}) + b_q \le f_{q+1}(\mathrm{x}) + b_{q+1},\ \forall q \in [Q-1],\ \forall \mathrm{x} \in K.$$

Task-1: joint quantile regression (JQR)

- Given: $(\tau_q)_{q\in[Q]} \subset (0,1)$ levels $\nearrow$, $\{(\mathrm{x}_n, y_n)\}_{n\in[N]}$ samples.
- Estimate jointly the τ_q-quantiles of $\mathbb{P}(Y|X = \mathrm{x})$[Sangnier et al., 2016].
- Objective:

$$\mathcal{L}(\mathrm{f},\mathrm{b}) = \frac{1}{N}\sum_{q\in[Q]}\sum_{n\in[N]} l_{\tau_q}\left(y_n - [f_q(\mathrm{x}_n) + b_q]\right) + \lambda_{\mathrm{b}}\|\mathrm{b}\|_2^2 + \lambda_f \sum_{q\in[Q]} \|f_q\|_k^2,$$

$$l_\tau(e) = \max(\tau e, (\tau - 1)e).$$

- Constraint (non-crossing): $K :=$ smallest rectangle containing $\{\mathrm{x}_n\}_{n\in[N]}$,

$$f_q(\mathrm{x}) + b_q \le f_{q+1}(\mathrm{x}) + b_{q+1},\ \forall q \in [Q-1],\ \forall \mathrm{x} \in K.$$

Constraints

function values (f_q) with interaction $(f_{q+1} - f_q)$, bias terms (b_q) with interaction $(b_q - b_{q+1})$.

Task-2: convoy localization, one vehicle ($Q = 1$)

- Given: noisy time-location samples $\{(t_n, x_n)\}_{n\in[N]} \subset \underbrace{[0, T]}_{=:\mathcal{T}} \times \mathbb{R}$.
- Goal: learn the (t, x) relation.
- Constraint: lower bound on speed (v_{min}).

Task-2: convoy localization, one vehicle ($Q = 1$)

- Given: noisy time-location samples $\{(t_n, x_n)\}_{n\in[N]} \subset \underbrace{[0, T]}_{=:\mathcal{T}} \times \mathbb{R}$.
- Goal: learn the (t, x) relation.
- Constraint: lower bound on speed ($v_{\min}$).
- Objective:

$$\begin{aligned}
&\min_{b \in \mathbb{R}, f \in \mathcal{H}_k} \left[\frac{1}{N} \sum_{n\in[N]} |x_n - (b + f(t_n))|^2 + \lambda \|f\|^2_{\mathcal{H}_k} \right] \\
&\text{s.t.} \\
&v_{\min} \le f'(t), \quad \forall t \in \mathcal{T}.
\end{aligned}$$

Task-2b: convoy localization, multiple vehicles ($Q \geq 1$)

- Data: $\left\{ (t_{q,n}, x_{q,n})_{n \in [N_q]} \right\}_{q \in [Q]} \subseteq \mathcal{T} \times \mathbb{R}$.
- Constraints: speed ($v_{\min}$), inter-vehicular distance ($d_{\min}$).
- Objective:

$$\min_{\substack{f_1,\ldots,f_Q \in \mathcal{H}_k, \\ b_1,\ldots,b_Q \in \mathbb{R}}} \frac{1}{Q} \sum_{q=1}^{Q} \left[\left(\frac{1}{N_q} \sum_{n=1}^{N_q} |x_{q,n} - (b_q + f_q(t_{q,n}))|^2 \right) + \lambda \|f_q\|_{\mathcal{H}_k}^2 \right]$$

s.t.

$$d_{\min} + b_{q+1} + f_{q+1}(t) \leq b_q + f_q(t), \forall q \in [Q-1], \; t \in \mathcal{T},$$

$$v_{\min} \leq f_q'(t), \qquad \forall q \in [Q], \; t \in \mathcal{T}.$$

Task-2b: convoy localization, multiple vehicles ($Q \geq 1$)

- Data: $\left\{(t_{q,n}, x_{q,n})_{n\in[N_q]}\right\}_{q\in[Q]} \subseteq \mathcal{T} \times \mathbb{R}$.
- Constraints: speed ($v_{\min}$), inter-vehicular distance ($d_{\min}$).
- Objective:

$$\min_{\substack{f_1,\ldots,f_Q \in \mathcal{H}_k,\\ b_1,\ldots,b_Q \in \mathbb{R}}} \frac{1}{Q}\sum_{q=1}^{Q}\left[\left(\frac{1}{N_q}\sum_{n=1}^{N_q}\left|x_{q,n} - (b_q + f_q(t_{q,n}))\right|^2\right) + \lambda\|f_q\|^2_{\mathcal{H}_k}\right]$$

s.t.

$$d_{\min} + b_{q+1} + f_{q+1}(t) \leq b_q + f_q(t), \forall q \in [Q-1],\ t \in \mathcal{T},$$
$$v_{\min} \leq f_q'(t), \qquad \forall q \in [Q],\ t \in \mathcal{T}.$$

Constraints

function values (f_q) and derivatives (f_q') with interaction ($f_q - f_{q+1}$), bias terms (b_q) with interaction ($b_{q+1} - b_q$).

Task-3: safety-critical control

- Trajectory of an underwater vehicle:

$$t \in \mathcal{T} := [0, 1] \mapsto [x(t); z(t)] \in \mathbb{R}^2.$$

Task-3: safety-critical control

- Trajectory of an underwater vehicle:

$$t \in \mathcal{T} := [0,1] \mapsto [x(t); z(t)] \in \mathbb{R}^2.$$

- Simplifying assumption: $x(0) = 0, \dot{x}(t) = 1\, \forall t \in \mathcal{T} \Rightarrow x(t) = t.$

Task-3: safety-critical control

- Trajectory of an underwater vehicle:

$$t \in \mathcal{T} := [0, 1] \mapsto [x(t); z(t)] \in \mathbb{R}^2.$$

- Simplifying assumption: $x(0) = 0, \dot{x}(t) = 1\, \forall t \in \mathcal{T} \Rightarrow x(t) = t$.
- Requirement: stay between the floor and the ceiling of the cavern

$$z(t) \in [z_{\text{low}}(t), z_{\text{up}}(t)] \;\; \forall t \in \mathcal{T}.$$

Task-3: safety-critical control

- Trajectory of an underwater vehicle:

$$t \in \mathcal{T} := [0,1] \mapsto [x(t); z(t)] \in \mathbb{R}^2.$$

- Simplifying assumption: $x(0) = 0, \dot{x}(t) = 1 \, \forall t \in \mathcal{T} \Rightarrow x(t) = t$.
- Requirement: stay between the floor and the ceiling of the cavern

$$z(t) \in [z_{\text{low}}(t), z_{\text{up}}(t)] \;\; \forall t \in \mathcal{T}.$$

- Initial condition: $z(0) = 0$ and $\dot{z}(0) = 0$.

Task-3: safety-critical control

- Trajectory of an underwater vehicle:

$$t \in \mathcal{T} := [0,1] \mapsto [x(t); z(t)] \in \mathbb{R}^2.$$

- Simplifying assumption: $x(0) = 0, \dot{x}(t) = 1 \, \forall t \in \mathcal{T} \Rightarrow x(t) = t$.
- Requirement: stay between the floor and the ceiling of the cavern

$$z(t) \in [z_{\text{low}}(t), z_{\text{up}}(t)] \;\; \forall t \in \mathcal{T}.$$

- Initial condition: $z(0) = 0$ and $\dot{z}(0) = 0$.
- Control task (LQ = linear dynamics & quadratic cost):

$$\begin{aligned} &\min_{u \in L^2(\mathcal{T},\mathbb{R})} \quad \int_{\mathcal{T}} |u(t)|^2 \mathrm{d}t \\ &\text{s.t.} \\ &z(0) = 0, \quad \dot{z}(0) = 0, \\ &\ddot{z}(t) = -\dot{z}(t) + u(t), \, \forall t \in \mathcal{T}, \\ &z_{\text{low}}(t) \leq z(t) \leq z_{\text{up}}(t), \, \forall \, t \in \mathcal{T}. \end{aligned}$$

Task-3: safety-critical control – continued

- With full state $\mathsf{f}(t) := [z(t); \dot{z}(t)] \in \mathbb{R}^2$

$$\dot{\mathsf{f}}(t) = \mathsf{A}\mathsf{f}(t) + \mathsf{B}u(t), \quad \mathsf{f}(0) = 0, \quad \mathsf{A} = \begin{bmatrix} 0 & 1 \\ 0 & -1 \end{bmatrix} \in \mathbb{R}^{2\times 2}, \quad \mathsf{B} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \in \mathbb{R}^2$$

Task-3: safety-critical control – continued

- With full state $\mathsf{f}(t) := [z(t); \dot{z}(t)] \in \mathbb{R}^2$

$$\dot{\mathsf{f}}(t) = \mathsf{A}\mathsf{f}(t) + \mathsf{B}u(t), \quad \mathsf{f}(0) = 0, \quad \mathsf{A} = \begin{bmatrix} 0 & 1 \\ 0 & -1 \end{bmatrix} \in \mathbb{R}^{2\times 2}, \quad \mathsf{B} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \in \mathbb{R}^2$$

- The controlled trajectories f belong to a $\mathbb{R}^2$-valued RKHS with kernel

$$k(s,t) := \int_0^{\min(s,t)} e^{(s-\tau)\mathsf{A}}\mathsf{B}\mathsf{B}^\top e^{(t-\tau)\mathsf{A}^\top} \mathrm{d}\tau, \quad s,t \in \mathcal{T},$$

and the task is

$$\begin{aligned} \min_{\mathsf{f}=[f_1;f_2]\in\mathcal{H}_k} \quad & \|\mathsf{f}\|_k^2 \\ \text{s.t.} \quad & \\ z_{\mathrm{low}}(t) \le f_1(t) \le z_{\mathrm{up}}(t),\ & \forall\, t \in \mathcal{T}. \end{aligned}$$

Task-3: safety-critical control – finished

- Assume for simplicity: z_{low} and z_{up} are piece-wise constant.
- Task:

$$\begin{aligned} \min_{\mathsf{f}=[f_1;f_2]\in\mathcal{H}_k} \quad & \|\mathsf{f}\|_k^2 \\ \text{s.t.} \quad & \\ z_{\text{low},m} \le f_1(t) \le z_{\text{up},m},\ & \forall\, t \in \mathcal{T}_m,\ \forall m \in [M]. \end{aligned}$$

Task-3: safety-critical control – finished

- Assume for simplicity: z_{low} and z_{up} are piece-wise constant.
- Task:

$$\begin{aligned} \min_{\mathrm{f}=[f_1;f_2]\in\mathcal{H}_k} \quad & \|\mathrm{f}\|_k^2 \\ \text{s.t.} \quad & \\ z_{\text{low},m} \le f_1(t) \le z_{\text{up},m}, \ & \forall\, t \in \mathcal{T}_m, \forall m \in [M]. \end{aligned}$$

Constraints

linear transformation of functions (f_1), with matrix-valued kernel.

Our task

$$(\bar{\mathrm{f}}, \bar{\mathrm{b}}) = \mathop{\arg\min}_{\substack{\mathrm{f}=(f_q)_{q\in[Q]} \in (\mathcal{H}_k)^Q, \\ \mathrm{b}=(b_q)_{q\in[Q]} \in \mathcal{B}, \\ (\mathrm{f},\mathrm{b}) \in C}} \mathcal{L}(\mathrm{f}, \mathrm{b}),$$

Our task

$$(\bar{\mathrm{f}}, \bar{\mathrm{b}}) = \mathop{\arg\min}_{\substack{\mathrm{f}=(f_q)_{q\in[Q]} \in (\mathcal{H}_k)^Q,\\ \mathrm{b}=(b_q)_{q\in[Q]} \in \mathcal{B},\\ (\mathrm{f},\mathrm{b}) \in C}} \mathcal{L}(\mathrm{f}, \mathrm{b}),$$

$$\mathcal{L}(\mathrm{f}, \mathrm{b}) = L\left(\mathrm{b}, \left(\mathrm{x}_n, y_n, (f_q(\mathrm{x}_n))_{q\in[Q]}\right)_{n\in[N]}\right) + \Omega\left((\|f_q\|_{\mathcal{H}_k})_{q\in[Q]}\right),$$

Our task

$$(\bar{\mathrm{f}}, \bar{\mathrm{b}}) = \underset{\substack{\mathrm{f}=(f_q)_{q\in[Q]} \in (\mathcal{H}_k)^Q,\\ \mathrm{b}=(b_q)_{q\in[Q]} \in \mathcal{B},\\ (\mathrm{f},\mathrm{b}) \in C}}{\arg\min} \mathcal{L}(\mathrm{f}, \mathrm{b}),$$

$$\mathcal{L}(\mathrm{f}, \mathrm{b}) = L\left(\mathrm{b}, \left(\mathrm{x}_n, y_n, (f_q(\mathrm{x}_n))_{q\in[Q]}\right)_{n\in[N]}\right) + \Omega\left((\|f_q\|_{\mathcal{H}_k})_{q\in[Q]}\right),$$

$$C = \{(\mathrm{f}, \mathrm{b}) \,|\, (\mathrm{b}_0 - \mathrm{U}\mathrm{b})_i \leq D_i(\mathrm{W}\mathrm{f} - \mathrm{f}_0)_i(\mathrm{x}), \quad \forall \mathrm{x} \in K_i, \forall i \in [I]\},$$

Our task

$$(\bar{\mathrm{f}}, \bar{\mathrm{b}}) = \operatorname*{arg\,min}_{\substack{\mathrm{f}=(f_q)_{q\in[Q]} \in (\mathcal{H}_k)^Q, \\ \mathrm{b}=(b_q)_{q\in[Q]} \in \mathcal{B}, \\ (\mathrm{f},\mathrm{b}) \in C}} \mathcal{L}(\mathrm{f}, \mathrm{b}),$$

$$\mathcal{L}(\mathrm{f}, \mathrm{b}) = L\left(\mathrm{b}, \left(\mathrm{x}_n, y_n, (f_q(\mathrm{x}_n))_{q\in[Q]}\right)_{n\in[N]}\right) + \Omega\left((\|f_q\|_{\mathcal{H}_k})_{q\in[Q]}\right),$$

$$C = \{(\mathrm{f}, \mathrm{b}) \,|\, (\mathrm{b}_0 - \mathrm{U}\mathrm{b})_i \le D_i(\mathrm{W}\mathrm{f} - \mathrm{f}_0)_i(\mathrm{x}), \quad \forall \mathrm{x} \in K_i, \forall i \in [I]\},$$

$$(\mathrm{W}\mathrm{f})_i = \sum_{q\in[Q]} W_{i,q} f_q,$$

Our task

$$(\bar{\mathrm{f}}, \bar{\mathrm{b}}) = \mathop{\arg\min}_{\substack{\mathrm{f}=(f_q)_{q\in[Q]} \in (\mathcal{H}_k)^Q, \\ \mathrm{b}=(b_q)_{q\in[Q]} \in \mathcal{B}, \\ (\mathrm{f},\mathrm{b}) \in C}} \mathcal{L}(\mathrm{f}, \mathrm{b}),$$

$$\mathcal{L}(\mathrm{f}, \mathrm{b}) = L\left(\mathrm{b}, \left(\mathrm{x}_n, y_n, (f_q(\mathrm{x}_n))_{q\in[Q]}\right)_{n\in[N]}\right) + \Omega\left((\|f_q\|_{\mathcal{H}_k})_{q\in[Q]}\right),$$

$$C = \{(\mathrm{f}, \mathrm{b}) \,|\, (\mathrm{b}_0 - \mathrm{U}\mathrm{b})_i \le D_i(\mathrm{W}\mathrm{f} - \mathrm{f}_0)_i(\mathrm{x}), \quad \forall \mathrm{x} \in K_i, \forall i \in [I]\},$$

$$(\mathrm{W}\mathrm{f})_i = \sum_{q\in[Q]} W_{i,q} f_q,$$

$$D_i = \sum_{j\in[n_{i,j}]} \gamma_{i,j} \partial^{\mathrm{r}_{i,j}},\ |\mathrm{r}_{i,j}| \le s,\ \gamma_{i,j} \in \mathbb{R},\ \partial^{\mathrm{r}} f(\mathrm{x}) = \frac{\partial^{|\mathrm{r}|} f(\mathrm{x})}{\partial^{r_1}_{x_1} \cdots \partial^{r_d}_{x_d}}.$$

Blanket assumptions

1. Domain $\mathcal{X} \subseteq \mathbb{R}^d$: open. Kernel $k \in \mathcal{C}^s(\mathcal{X} \times \mathcal{X})$.
2. $K_i \subset \mathcal{X}$: compact, $\forall i$.
3. $\mathrm{f}_{0,i} \in \mathcal{H}_k$ for $\forall i$.
4. Bias domain $\mathcal{B} \subseteq \mathbb{R}^Q$: convex.
5. Loss L restricted to $\mathcal{B}$: strictly convex in b.
6. Regularizer Ω: strictly increasing in each of its argument.

Our strenghtened SOC-constrained formulation

$$(\mathsf{f}_{\boldsymbol{\eta}}, \mathsf{b}_{\boldsymbol{\eta}}) = \underset{\mathsf{f} \in (\mathcal{H}_k)^Q,\, \mathsf{b} \in \mathcal{B}}{\arg\min} \quad \mathcal{L}(\mathsf{f}, \mathsf{b}) \qquad (\mathcal{P}_{\boldsymbol{\eta}})$$

$$\text{s.t.}$$

$$(\mathsf{b}_0 - \mathsf{U}\mathsf{b})_i + \eta_i \|(\mathsf{W}\mathsf{f} - \mathsf{f}_0)_i\|_{\mathcal{H}_k} \leq \min_{m \in [M_i]} D_i(\mathsf{W}\mathsf{f} - \mathsf{f}_0)_i\left(\tilde{\mathsf{x}}_{i,m}\right), \ \forall i \in [I], \qquad (\mathcal{C}_{\boldsymbol{\eta}})$$

where

- $\{\tilde{\mathsf{x}}_{i,m}\}_{m \in [M_i]}$: a δ_i-net of K_i in $\|\cdot\|_{\mathcal{X}}$,
- $\eta_i = \sup_{m \in [M_i], \mathsf{u} \in \mathbb{B}_{\|\cdot\|_{\mathcal{X}}}(0,1)} \|D_{i,\mathsf{x}} k(\tilde{\mathsf{x}}_{i,m}, \cdot) - D_{i,\mathsf{x}} k(\tilde{\mathsf{x}}_{i,m} + \delta_i \mathsf{u}, \cdot)\|_{\mathcal{H}_k}$.

Theorem

- Minimal values: v_{disc} = value of $(\mathcal{P}_{\boldsymbol{\eta}})$ with '$\boldsymbol{\eta} = 0$', $\bar{v} = \mathcal{L}\left(\bar{\mathrm{f}}, \bar{\mathrm{b}}\right)$, $v_{\boldsymbol{\eta}} = \mathcal{L}\left(\mathrm{f}_{\boldsymbol{\eta}}, \mathrm{b}_{\boldsymbol{\eta}}\right)$.
- Let $\mathrm{f}_{\boldsymbol{\eta}} = \left(f_{\boldsymbol{\eta},q}\right)_{q\in[Q]}$.

Theorem

- Minimal values: v_{disc} = value of $(\mathcal{P}_{\boldsymbol{\eta}})$ with '$\boldsymbol{\eta} = 0$', $\bar{v} = \mathcal{L}\left(\bar{\mathrm{f}}, \bar{\mathrm{b}}\right)$, $v_{\boldsymbol{\eta}} = \mathcal{L}\left(\mathrm{f}_{\boldsymbol{\eta}}, \mathrm{b}_{\boldsymbol{\eta}}\right)$.
- Let $\mathrm{f}_{\boldsymbol{\eta}} = \left(f_{\boldsymbol{\eta},q}\right)_{q \in [Q]}$.

Then,

- (i) Tightening: any (f, b) satisfying $(\mathcal{C}_{\boldsymbol{\eta}})$ also satisfies $(\mathcal{C})$, hence

$$v_{\text{disc}} \leq \bar{v} \leq v_{\boldsymbol{\eta}}.$$

Theorem

- Minimal values: v_{disc} = value of $(\mathcal{P}_{\boldsymbol{\eta}})$ with '$\boldsymbol{\eta} = 0$', $\bar{v} = \mathcal{L}\left(\bar{\mathrm{f}}, \bar{\mathrm{b}}\right)$, $v_{\boldsymbol{\eta}} = \mathcal{L}\left(\mathrm{f}_{\boldsymbol{\eta}}, \mathrm{b}_{\boldsymbol{\eta}}\right)$.
- Let $\mathrm{f}_{\boldsymbol{\eta}} = (f_{\boldsymbol{\eta},q})_{q\in[Q]}$.

Then,

- (i) Tightening: any (f, b) satisfying $(\mathcal{C}_{\boldsymbol{\eta}})$ also satisfies $(\mathcal{C})$, hence

$$v_{\text{disc}} \leq \bar{v} \leq v_{\boldsymbol{\eta}}.$$

- (ii) Representer theorem: For $\forall q \in [Q]$, $\exists \tilde{a}_{i,0,q}, \tilde{a}_{i,m,q}, a_{n,q} \in \mathbb{R}$ s.t.

$$f_{\boldsymbol{\eta},q} = \sum_{i\in[I]} \left[\tilde{a}_{i,0,q} f_{0,i} + \sum_{m\in[M_i]} \tilde{a}_{i,m,q} D_{i,\mathrm{x}} k\left(\tilde{\mathrm{x}}_{i,m}, \cdot\right)\right] + \sum_{n\in[N]} a_{n,q} k(\mathrm{x}_n, \cdot).$$

Theorem – continued

- (iii) Performance guarantee: if $\mathcal{L}$ is $(\mu_{f_q}, \mu_{\mathrm{b}})$-strongly convex w.r.t. (f_q, b) for any $q \in [Q]$, then

$$\|f_{\boldsymbol{\eta},q} - \bar{f}_q\|_{\mathcal{H}_k} \leq \sqrt{\frac{2(v_{\boldsymbol{\eta}} - v_{\mathrm{disc}})}{\mu_{f_q}}}, \quad \|\mathrm{b}_{\boldsymbol{\eta}} - \bar{\mathrm{b}}\|_2 \leq \sqrt{\frac{2(v_{\boldsymbol{\eta}} - v_{\mathrm{disc}})}{\mu_{\mathrm{b}}}}.$$

Theorem – continued

- (iii) Performance guarantee: if $\mathcal{L}$ is $(\mu_{f_q}, \mu_{\mathrm{b}})$-strongly convex w.r.t. (f_q, b) for any $q \in [Q]$, then

$$\|f_{\boldsymbol{\eta},q} - \bar{f}_q\|_{\mathcal{H}_k} \leq \sqrt{\frac{2(v_{\boldsymbol{\eta}} - v_{\mathrm{disc}})}{\mu_{f_q}}}, \quad \|\mathrm{b}_{\boldsymbol{\eta}} - \bar{\mathrm{b}}\|_2 \leq \sqrt{\frac{2(v_{\boldsymbol{\eta}} - v_{\mathrm{disc}})}{\mu_{\mathrm{b}}}}.$$

If in addition U is surjective, $\mathcal{B} = \mathbb{R}^Q$, and $\mathcal{L}(\bar{\mathrm{f}}, \cdot)$ is L_b–Lipschitz continuous on $\mathbb{B}_{\|\cdot\|_2}\left(\bar{\mathrm{b}}, c_f\|\boldsymbol{\eta}\|_\infty\right)$ where $c_f \;=\; \sqrt{d}\left\|\left(\mathrm{U}^T\mathrm{U}\right)^{-1}\mathrm{U}^T\right\| \max_{i\in[I]} \left\|(\mathrm{W}\bar{\mathrm{f}} - \mathrm{f}_0)_i\right\|_{\mathcal{H}_k}$, then

$$\|f_{\boldsymbol{\eta},q} - \bar{f}_q\|_{\mathcal{H}_k} \leq \sqrt{\frac{2L_b c_f\|\boldsymbol{\eta}\|_\infty}{\mu_{f_q}}}, \|\mathrm{b}_{\boldsymbol{\eta}} - \bar{\mathrm{b}}\|_2 \leq \sqrt{\frac{2L_b c_f\|\boldsymbol{\eta}\|_\infty}{\mu_{\mathrm{b}}}}.$$

Theorem – continued

- (iii) Performance guarantee: if $\mathcal{L}$ is $(\mu_{f_q}, \mu_{\mathrm{b}})$-strongly convex w.r.t. (f_q, b) for any $q \in [Q]$, then

$$\|f_{\boldsymbol{\eta},q} - \bar{f}_q\|_{\mathcal{H}_k} \leq \sqrt{\frac{2(v_{\boldsymbol{\eta}} - v_{\mathrm{disc}})}{\mu_{f_q}}}, \quad \|\mathrm{b}_{\boldsymbol{\eta}} - \bar{\mathrm{b}}\|_2 \leq \sqrt{\frac{2(v_{\boldsymbol{\eta}} - v_{\mathrm{disc}})}{\mu_{\mathrm{b}}}}.$$

If in addition U is surjective, $\mathcal{B} = \mathbb{R}^Q$, and $\mathcal{L}(\bar{\mathrm{f}}, \cdot)$ is L_b–Lipschitz continuous on $\mathbb{B}_{\|\cdot\|_2}\left(\bar{\mathrm{b}}, c_f\|\boldsymbol{\eta}\|_\infty\right)$ where $c_f = \sqrt{d}\left\|\left(\mathrm{U}^T\mathrm{U}\right)^{-1}\mathrm{U}^T\right\| \max_{i\in[I]} \|(\mathrm{W}\bar{\mathrm{f}} - \mathrm{f}_0)_i\|_{\mathcal{H}_k}$, then

$$\|f_{\boldsymbol{\eta},q} - \bar{f}_q\|_{\mathcal{H}_k} \leq \sqrt{\frac{2L_b c_f\|\boldsymbol{\eta}\|_\infty}{\mu_{f_q}}}, \|\mathrm{b}_{\boldsymbol{\eta}} - \bar{\mathrm{b}}\|_2 \leq \sqrt{\frac{2L_b c_f\|\boldsymbol{\eta}\|_\infty}{\mu_{\mathrm{b}}}}.$$

1st bound: computable. 2nd: Larger $M_i \Rightarrow$ smaller $\delta_i \Rightarrow$ smaller η_i $\Rightarrow$ tighter bound.

Tightening idea

Let $s = 0$, $l = 1$. Recall constraint $(\mathcal{C})$:

$$\left\{(\mathrm{f}, \mathrm{b}) \mid \underbrace{(b_0 - \mathrm{Ub})}_{\beta} \leq \underbrace{\underbrace{(\mathrm{Wf} - f_0)}_{\phi}(\mathrm{x})}_{\langle \phi, k(\mathrm{x}, \cdot) \rangle_{\mathcal{H}_k}}, \quad \forall \mathrm{x} \in K\right\}$$

Tightening idea

Let $s = 0$, $l = 1$. Recall constraint $(\mathcal{C})$:

$$\Big\{(\mathrm{f},\mathrm{b}) \,|\, \underbrace{(b_0 - \mathrm{Ub})}_{\beta} \le \underbrace{\underbrace{(\mathrm{Wf} - f_0)}_{\phi}(\mathrm{x})}_{\langle \phi, k(\mathrm{x},\cdot)\rangle_{\mathcal{H}_k}}, \quad \forall \mathrm{x} \in K\Big\}, \text{i.e.}$$

$$\boxed{\Phi(K)} := \{k(\mathrm{x},\cdot) : \mathrm{x} \in K\} \boxed{\subseteq H^+_{\phi,\beta}} := \Big\{g \in \mathcal{H}_k \,|\, \beta \le \langle \phi, g\rangle_{\mathcal{H}_k}\Big\}$$

Tightening idea

Let $s = 0$, $l = 1$. Recall constraint $(\mathcal{C})$:

$$\Big\{(\mathrm{f}, \mathrm{b}) \mid \underbrace{(b_0 - \mathrm{Ub})}_{\beta} \leq \underbrace{\underbrace{(\mathrm{Wf} - f_0)}_{\phi}(\mathrm{x})}_{\langle \phi, k(\mathrm{x}, \cdot)\rangle_{\mathcal{H}_k}}, \quad \forall \mathrm{x} \in K\Big\}, \text{ i.e.}$$

$$\boxed{\Phi(K)} := \{k(\mathrm{x}, \cdot) : \mathrm{x} \in K\} \boxed{\subseteq H^{+}_{\phi,\beta}} := \Big\{g \in \mathcal{H}_k \mid \beta \leq \langle \phi, g\rangle_{\mathcal{H}_k}\Big\}$$

- $(\mathcal{C}_{\boldsymbol{\eta}})$ means: covering of $\Phi(K)$ by balls with η-radius centered at the $k(\tilde{\mathrm{x}}_m, \cdot)$ is in the halfspace $H^{+}_{\phi,\beta}$; hence it is tightening.

Tightening idea

Let $s = 0$, $l = 1$. Recall constraint $(\mathcal{C})$:

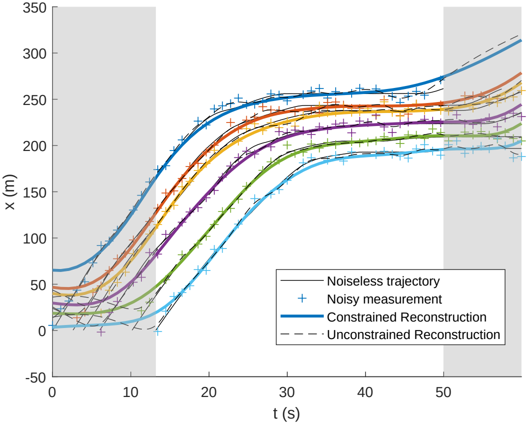
$$\Big\{(\mathrm{f},\mathrm{b}) \,|\, \underbrace{(b_0 - \mathrm{Ub})}_{\beta} \le \underbrace{\underbrace{({\color{red}\mathrm{Wf}} - f_0)}_{\phi}(\mathrm{x})}_{\langle \phi, k(\mathrm{x},\cdot)\rangle_{\mathcal{H}_k}}, \quad \forall \mathrm{x} \in K\Big\}, \text{i.e.}$$

$$\boxed{\Phi(K)} := \{k(\mathrm{x},\cdot) : \mathrm{x} \in K\} \boxed{\subseteq H^+_{\phi,\beta}} := \Big\{g \in \mathcal{H}_k \,|\, \beta \le \langle \phi, g\rangle_{\mathcal{H}_k}\Big\}$$

- $(\mathcal{C}_{\boldsymbol{\eta}})$ means: covering of $\Phi(K)$ by balls with η-radius centered at the $k(\tilde{\mathrm{x}}_m, \cdot)$ is in the halfspace $H^+_{\phi,\beta}$; hence it is tightening.
- η is obtained as the minimal radius.

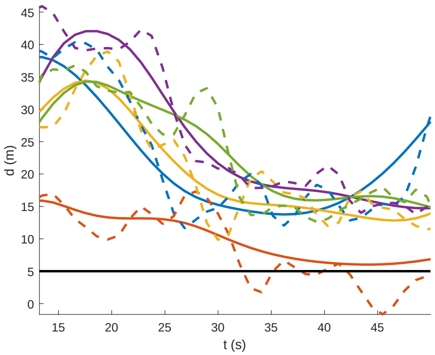
Demo (task-1): convoy localization with traffic jam

Setting: $Q = 6$, $d_{\min} = 5m$, $v_{\min} = 0$.



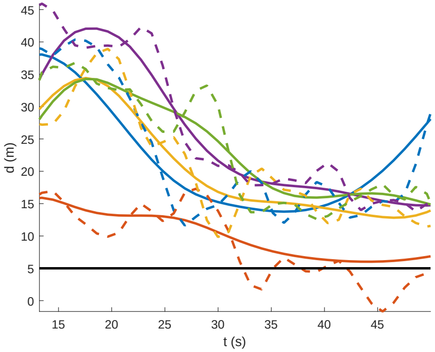
Demo (task-1): continued

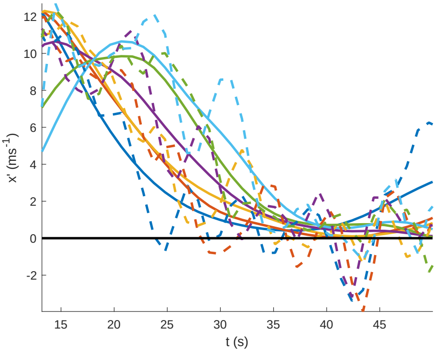
Pairwise distances: $t \mapsto f_q(t) - f_{q+1}(t)$



Demo (task-1): continued

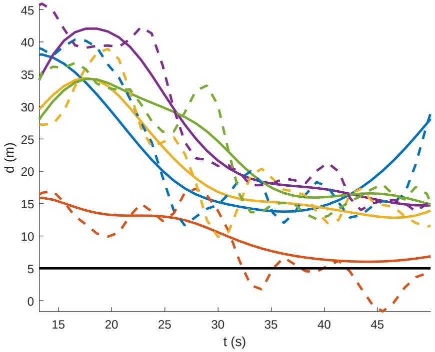
Pairwise distances: $t \mapsto f_q(t) - f_{q+1}(t)$ Speed: $t \mapsto f'_q(t)$

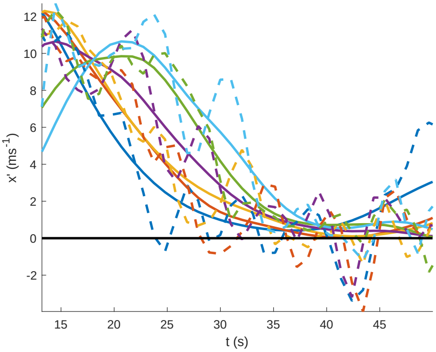




Demo (task-1): continued

Pairwise distances: $t \mapsto f_q(t) - f_{q+1}(t)$ Speed: $t \mapsto f_q'(t)$



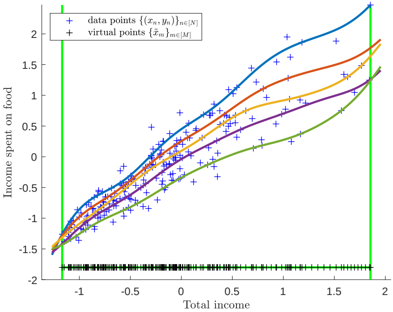


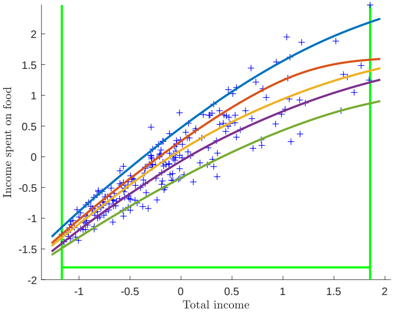
Shape constraints: especially relevant in noisy situations.

Demo (task-2): joint quantile regression

Economics :

- x: annual household income, y: food expenditure. $d = 1$, $N = 235$.
- Engel's law $\Rightarrow \nearrow$, concave.
- Demo: $\tau_q \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$.
- Left: non-crossing, $\nearrow$. Right: non-crossing, $\nearrow$, concave.

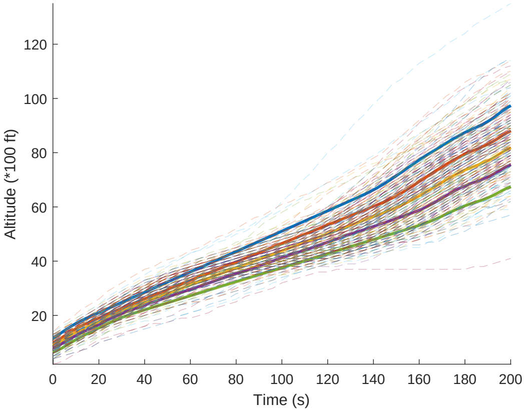




Demo (task-2): joint quantile regression

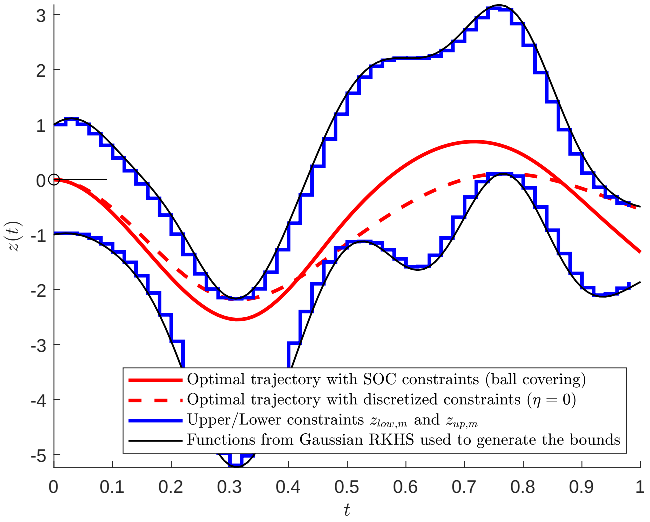
Analysis of aircraft trajectories, ENAC:

- y: radar-measured altitude of aircrafts flying between two cities (Paris & Toulouse); x: time. $d = 1$, $N = 15657$.
- Demo: $\tau_q \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$.
- Constraint: non-crossing, $\nearrow$ (takeoff).



Demo (task-3): control of underwater vehicle

Vs discretization-based approach (which might crash):



Summary

- Focus: hard affine shape constraints on derivatives & RKHS.
- Proposed framework: SOC-based tightening.
- Applications:
  - convoy localization,
  - joint quantile regression: economics, aircraft trajectories,
  - safety-critical control.

References & acknowledgements

Details:

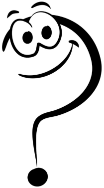
- Transportation systems [Aubin-Frankowski et al., 2020],
- Control aspect [Aubin-Frankowski, 2020],
- Method [Aubin-Frankowski and Szabó, 2020], extension [under preparation to JMLR]. Code @ GitHub.

Acknowledgements: ZSz benefited from the support of the Europlace Institute of Finance and that of the Chair Stress Test, RISK Management and Financial Steering, led by the French École Polytechnique and its Foundation and sponsored by BNP Paribas.

References & acknowledgements

Details:

- Transportation systems [Aubin-Frankowski et al., 2020],
- Control aspect [Aubin-Frankowski, 2020],
- Method [Aubin-Frankowski and Szabó, 2020], extension [under preparation to JMLR]. Code @ GitHub.



Acknowledgements: ZSz benefited from the support of the Europlace Institute of Finance and that of the Chair Stress Test, RISK Management and Financial Steering, led by the French École Polytechnique and its Foundation and sponsored by BNP Paribas.

INSTITUT Louis Bachelier

Stress Test
RISK Management and Financial Steering

Aït-Sahalia, Y. and Duarte, J. (2003).
Nonparametric option pricing under shape restrictions.
Journal of Econometrics, 116(1-2):9–47.

Aubin-Frankowski, P.-C. (2020).
Linearly-constrained linear quadratic regulator from the viewpoint of kernel methods.
Technical report.
(`https://arxiv.org/abs/2011.02196`).

Aubin-Frankowski, P.-C., Petit, N., and Szabó, Z. (2020).
Kernel regression for vehicle trajectory reconstruction under speed and inter-vehicular distance constraints.
In *IFAC World Congress (IFAC WC)*, Berlin, Germany.

Aubin-Frankowski, P.-C. and Szabó, Z. (2020).
Hard shape-constrained kernel machines.
In *Advances in Neural Information Processing Systems (NeurIPS)*.

Bach, F. and Jordan, M. (2002).
Kernel independent component analysis.
Journal of Machine Learning Research, 3:1–48.

Bai, L., Cui, L., Rossi, L., Xu, L., Bai, X., and Hancock, E. (2020).
Local-global nested graph kernels using nested complexity traces.
Pattern Recognition Letters, 134:87–95.

Balabdaoui, F., Durot, C., and Jankowski, H. (2019).
Least squares estimation in the monotone single index model.
Bernoulli, 25(4B):3276–3310.

Blanchard, G., Lee, G., and Scott, C. (2011).
Generalizing from several related classification tasks to a new unlabeled sample.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 2178–2186.

Blundell, R., Horowitz, J. L., and Parey, M. (2012).
Measuring the price responsiveness of gasoline demand: economic shape restrictions and nonparametric demand estimation.
Quantitative Economics, 3:29–51.

Borgwardt, K. M. and Kriegel, H.-P. (2005).
Shortest-path kernels on graphs.
In *International Conference on Data Mining (ICDM)*, pages 74–81.

Carmeli, C., Vito, E. D., and Toigo, A. (2006).
Vector valued reproducing kernel Hilbert spaces of integrable functions and Mercer theorem.
Analysis and Applications, 4:377–408.

Chen, Y. and Samworth, R. J. (2016).
Generalized additive and index models with shape constraints.
Journal of the Royal Statistical Society – Statistical Methodology, Series B, 78(4):729–754.

Chetverikov, D., Santos, A., and Shaikh, A. M. (2018).
The econometrics of shape restrictions.
Annual Review of Economics, 10(1):31–63.

Collins, M. and Duffy, N. (2001).
Convolution kernels for natural language.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 625–632.

Cuturi, M. (2011).
Fast global alignment kernels.
In *International Conference on Machine Learning (ICML)*, pages 929–936.

Cuturi, M., Fukumizu, K., and Vert, J.-P. (2005).
Semigroup kernels on measures.
Journal of Machine Learning Research, 6:1169–1198.

Cuturi, M. and Vert, J.-P. (2005).
The context-tree kernel for strings.

Neural Networks, 18(8):1111–1123.

Cuturi, M., Vert, J.-P., Birkenes, O., and Matsui, T. (2007).
A kernel for time series based on global alignments.
In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 413–416.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. (2008).
Kernel measures of conditional dependence.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 498–496.

Gärtner, T., Flach, P., Kowalczyk, A., and Smola, A. (2002).
Multi-instance kernels.
In *International Conference on Machine Learning (ICML)*, pages 179–186.

Gärtner, T., Flach, P., and Wrobel, S. (2003).
On graph kernels: Hardness results and efficient alternatives.
Learning Theory and Kernel Machines, pages 129–143.

Gretton, A. (2015).
A simpler condition for consistency of a kernel independence test.
Technical report, University College London.
(http://arxiv.org/abs/1501.06103).

Guevara, J., Hirata, R., and Canu, S. (2017).
Cross product kernels for fuzzy set similarity.
In *International Conference on Fuzzy Systems (FUZZ-IEEE)*, pages 1–6.

Guntuboyina, A. and Sen, B. (2018).
Nonparametric shape-restricted regression.
Statistical Science, 33(4):568–594.

Haussler, D. (1999).
Convolution kernels on discrete structures.
Technical report, University of California at Santa Cruz.
(http://cbse.soe.ucsc.edu/sites/default/files/convolutions.pdf).

Hu, J., Kapoor, M., Zhang, W., Hamilton, S. R., and Coombes, K. R. (2005).
Analysis of dose-response effects on gene expression data with comparison of two microarray platforms.
Bioinformatics, 21(17):3524–3529.

Jaakkola, T. S. and Haussler, D. (1999).
Exploiting generative models in discriminative classifiers.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 487–493.

Jebara, T., Kondor, R., and Howard, A. (2004).
Probability product kernels.
Journal of Machine Learning Research, 5:819–844.

Jiao, Y. and Vert, J.-P. (2016).
The Kendall and Mallows kernels for permutations.
In *International Conference on Machine Learning (ICML)*, volume 37, pages 2982–2990.

Johnson, A. L. and Jiang, D. R. (2018).
Shape constraints in economics and operations research.
Statistical Science, 33(4):527–546.

Kashima, H. and Koyanagi, T. (2002).
Kernels for semi-structured data.
In *International Conference on Machine Learning (ICML)*, pages 291–298.

Kashima, H., Tsuda, K., and Inokuchi, A. (2003).
Marginalized kernels between labeled graphs.
In *International Conference on Machine Learning (ICML)*, pages 321–328.

Keshavarz, A., Wang, Y., and Boyd, S. (2011).
Imputing a convex objective function.
In *IEEE Multi-Conference on Systems and Control*, pages 613–619.

Király, F. J. and Oberhauser, H. (2019).
Kernels for sequentially ordered data.

Journal of Machine Learning Research, 20:1–45.

Kondor, R. and Pan, H. (2016).
The multiscale Laplacian graph kernel.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 2982–2990.

Kondor, R. I. and Lafferty, J. (2002).
Diffusion kernels on graphs and other discrete input.
In *International Conference on Machine Learning (ICML)*, pages 315–322.

Kuang, R., Ie, E., Wang, K., Wang, K., Siddiqi, M., Freund, Y., and Leslie, C. (2004).
Profile-based string kernels for remote homology detection and motif extraction.
Journal of Bioinformatics and Computational Biology, 13(4):527–550.

Leslie, C., Eskin, E., and Noble, W. S. (2002).

The spectrum kernel: A string kernel for SVM protein classification.
Biocomputing, pages 564–575.

Leslie, C. and Kuang, R. (2004).
Fast string kernels using inexact matching for protein sequences.
Journal of Machine Learning Research, 5:1435–1455.

Lewbel, A. (2010).
Shape-invariant demand functions.
The Review of Economics and Statistics, 92(3):549–556.

Li, Q. and Racine, J. S. (2007).
Nonparametric Econometrics.
Princeton University Press.

Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., and Watkins, C. (2002).
Text classification using string kernels.
Journal of Machine Learning Research, 2:419–444.

Luss, R., Rossett, S., and Shahar, M. (2012).
Efficient regularized isotonic regression with application to gene-gene interaction search.
Annals of Applied Statistics, 6(1):253–283.

Matzkin, R. L. (1991).
Semiparametric estimation of monotone and concave utility functions for polychotomous choice models.
Econometrica, 59(5):1315–1327.

Micchelli, C. and Pontil, M. (2005).
On learning vector-valued functions.
Neural Computation, 17:177–204.

Micchelli, C., Xu, Y., and Zhang, H. (2006).
Universal kernels.
Journal of Machine Learning Research, 7:2651–2667.

Pedrick, G. (1957).

Theory of reproducing kernels for Hilbert spaces of vector valued functions.
Technical report.

Rüping, S. (2001).
SVM kernels for time series analysis.
Technical report, University of Dortmund.
(http://www.stefan-rueping.de/publications/rueping-2001-a.pdf).

Saigo, H., Vert, J.-P., Ueda, N., and Akutsu, T. (2004).
Protein homology detection using string alignment kernels.
Bioinformatics, 20(11):1682–1689.

Sangnier, M., Fercoq, O., and d'Alché Buc, F. (2016).
Joint quantile regression in vector-valued RKHSs.
Advances in Neural Information Processing Systems (NIPS), pages 3693–3701.

Seeger, M. (2002).
Covariance kernels from Bayesian generative models.

In *Advances in Neural Information Processing Systems (NIPS)*, pages 905–912.

Shapiro, A., Dentcheva, D., and Ruszczynski, A. (2014).
Lectures on Stochastic Programming: Modeling and Theory.
SIAM - Society for Industrial and Applied Mathematics.

Shervashidze, N., Vishwanathan, S. V. N., Petri, T., Mehlhorn, K., and Borgwardt, K. M. (2009).
Efficient graphlet kernels for large graph comparison.
In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 488–495.

Shi, X., Shum, M., and Song, W. (2018).
Estimating semi-parametric panel multinomial choice models using cyclic monotonicity.
Econometrica, 86(2):737–761.

Simchi-Levi, D., Chen, X., and Bramel, J. (2014).
The Logic of Logistics: Theory, Algorithms, and Applications for Logistics Management.

Springer.

Simon-Gabriel, C.-J. and Schölkopf, B. (2018).
Kernel distribution embeddings: Universal kernels, characteristic kernels and kernel metrics on distributions.
Journal of Machine Learning Research, 19(44):1–29.

Sriperumbudur, B., Fukumizu, K., and Lanckriet, G. (2011).
Universality, characteristic kernels and RKHS embedding of measures.
Journal of Machine Learning Research, 12:2389–2410.

Sriperumbudur, B., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. (2010).
Hilbert space embeddings and metrics on probability measures.
Journal of Machine Learning Research, 11:1517–1561.

Steinwart, I. (2001).
On the influence of the kernel on the consistency of support vector machines.

Journal of Machine Learning Research, 6(3):67–93.

Szabó, Z. and Sriperumbudur, B. K. (2018).
Characteristic and universal tensor product kernels.
Journal of Machine Learning Research, 18(233):1–29.

Topkis, D. M. (1998).
Supermodularity and complementarity.
Princeton University Press.

Tsuda, K., Kin, T., and Asai, K. (2002).
Marginalized kernels for biological sequences.
Bioinformatics, 18:268–275.

Varian, H. R. (1984).
The nonparametric approach to production analysis.
Econometrica, 52(3):579–597.

Vishwanathan, S. N., Schraudolph, N., Kondor, R., and Borgwardt, K. (2010).
Graph kernels.

Journal of Machine Learning Research, 11:1201–1242.

Watkins, C. (1999).
Dynamic alignment kernels.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 39–50.

Yagi, D., Chen, Y., Johnson, A. L., and Kuosmanen, T. (2020).
Shape-constrained kernel-weighted least squares: Estimating production functions for Chilean manufacturing industries.
Journal of Business & Economic Statistics, 38(1):43–54.