Kernel Regression with Hard Shape Constraints

Pierre-Cyril Aubin-Frankowski, Zoltán Szabó

PhD student (École des Ponts ParisTech),
CAS - Centre Automatique et Systèmes, MINES ParisTech

https://pcaubin.github.io/

SMAI-MODE, September 2020

Problem definition

We are are interested in kernel regression with several pointwise constraints over derivatives. Given samples $(x_n, y_n)_{n\in[N]} \in (\mathcal{X} \times \mathbb{R})^N$, a *loss* $L : (\mathcal{X} \times \mathbb{R} \times \mathbb{R})^N \to \mathbb{R}$, a *regularizer* $\Omega : \mathbb{R}_+ \to \mathbb{R}$, for $x \in \mathcal{X} \subset \mathbb{R}^d$, $f \in \mathcal{C}^s(\mathcal{X}, \mathbb{R})$,

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

where $\mathcal{F}_k$ is a Hilbert space of real-valued functions (RKHS) over $\mathcal{X}$, D_i is a differential operator ($D = \sum_j \gamma_j \partial^{r_j}$), $b_i \in \mathbb{R}$ is a bound, $\mathcal{K}_i$ is a compact set (e.g. $[0, T]$, $[0, 1]^d$).

For non-finite $\mathcal{K}$, we have an infinite number of constraints!
$\hookrightarrow$ No representer theorem!

How can we make the optimization problem computationally tractable?

Shape constraints, ex:

Shape constraints = priors on the form of the solution of the problem

$\hookrightarrow$ compensates lack of samples or excessive noise

$\hookrightarrow$ incorporates physical constraints

They are crucial if the output model is then used as an input for safety-critical tasks (e.g. path-planning) or for theoretical analysis (e.g. density estimation). Many shape constraints are defined pointwise ($0 \leq Df(x)$):

- Statistics [Koenker, 2005]: nonnegative densities, noncrossing quantiles
- Economics [Matzkin, 1991]: increasing and concave utility functions
- Control/Path-planning [Egerstedt, 2009]: state and control constraints
- Supply chain [Simchi-Levi et al., 2014], Pricing models: supermodularity

Dealing with an infinite number of constraints: an overview

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

Relaxing

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

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

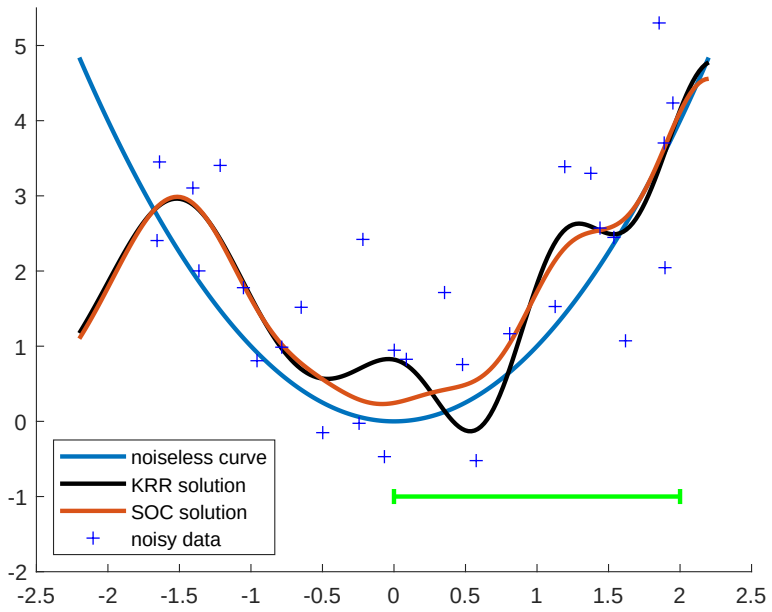
Tightening

- Replace $\mathcal{F}_k$ by algebraic subclass of functions satisfying the constraints
  $\hookrightarrow$ hard to stack constraints, $\Phi(x)^\top A \Phi(x)$, Sum-Of-Squares [Hall, 2018]

- Use only spaces $\mathcal{F}_k$ s.t. constraints have a "simple" writing, e.g. splines
  $\hookrightarrow$ highly restricted functions classes [Papp and Alizadeh, 2014]

- **Our solution:** discretize $\mathcal{K}_i$ but replace b_i using RKHS geometry

Example: 1D monotonic kernel ridge regression (KRR)



$$\min_{f \in \mathcal{F}_k} \frac{1}{N} \sum_{n=1}^{N} |y_n - f(x_n)|^2 + \lambda \|f\|_{\mathcal{F}_k}^2$$

$$\text{s.t. } 0 \leq f'(x), \forall x \in [0, 2]$$

Unconstrained KRR

vs

Second-Order Cone (SOC) constrained

SOC comes from adding a buffer to a discretization (interior solution)

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

How to choose η_m?

Reproducing kernel Hilbert spaces (RKHS) in one slide

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

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

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

k is s.t. $\mathbf{G} = [k(x_i, x_j)]_{i,j=1}^n \succcurlyeq 0$ and $\mathcal{F}_k := \overline{\mathrm{span}(\{k_x(\cdot)\}_{x \in \mathcal{X}})}$

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

- There is a one-to-one correspondence between kernels k and RKHSs $(\mathcal{F}_k, \langle \cdot, \cdot \rangle_{\mathcal{F}_k})$. Changing $\mathcal{X}$ or $\langle \cdot, \cdot \rangle_{\mathcal{F}_k}$ changes the kernel k.
- if $\mathcal{X} \subset \mathbb{R}^d$ is contained in closure of its interior, $k \in \mathcal{C}^{s,s}(\mathcal{X} \times \mathcal{X}, \mathbb{R})$, D a differential operator of order at most s, then kernel trick for derivatives:

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

Back to Second-Order Cone Constraints

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

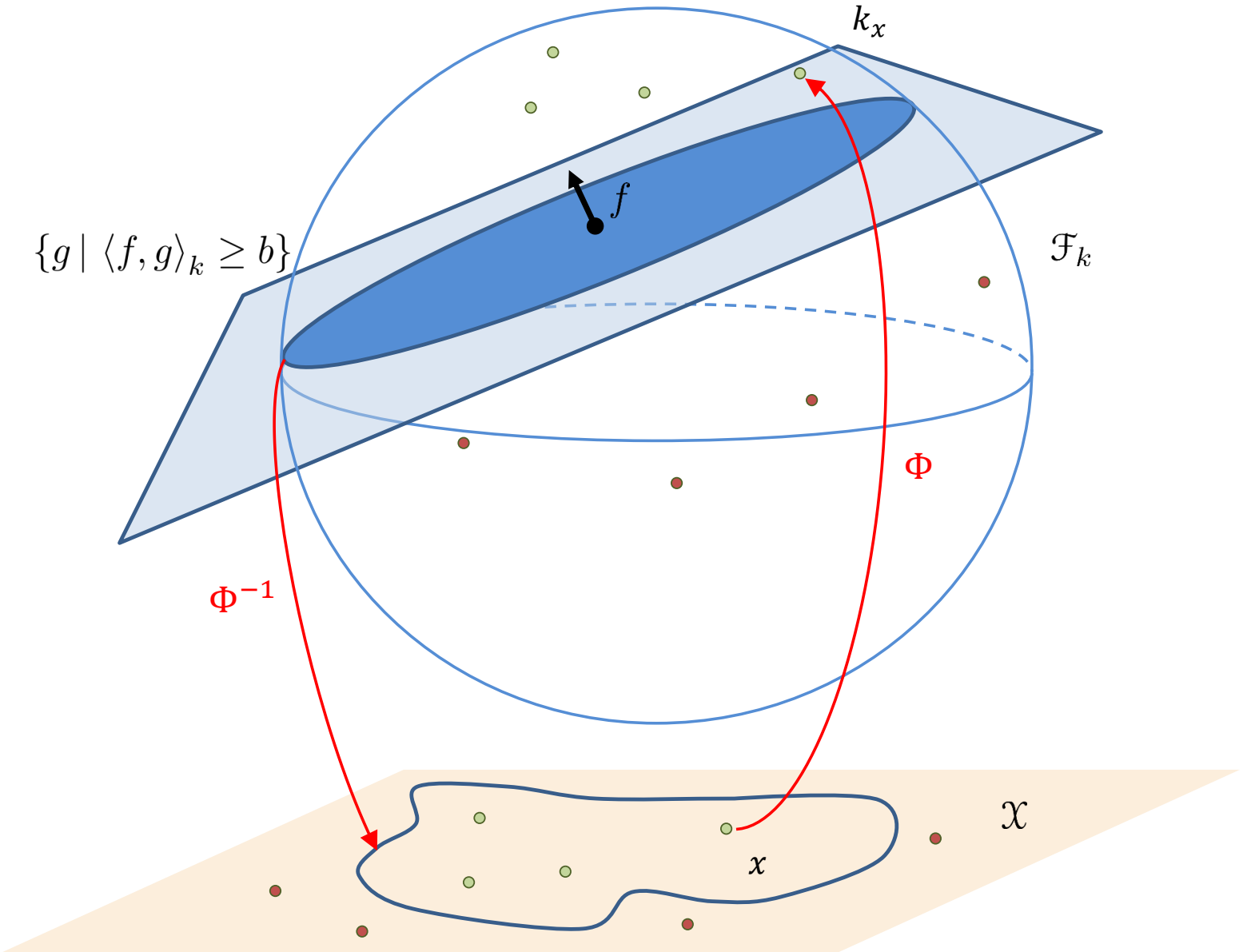
$$
\begin{aligned}
Df(x) &= Df(\tilde{x}_m) + \langle f(\cdot), D_x k(x, \cdot) - D_x k(\tilde{x}_m, \cdot) \rangle_k \\
Df(x) &\geq Df(\tilde{x}_m) - \|f(\cdot)\|_k \|D_x k(x, \cdot) - D_x k(\tilde{x}_m, \cdot)\|_k \\
Df(x) &\geq Df(\tilde{x}_m) - \|f(\cdot)\|_k \underbrace{\sup_{\{x \,|\, \|x - \tilde{x}_m\|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

For smooth kernels, $\delta \to 0$ gives $\eta_m(\delta) \to 0$.

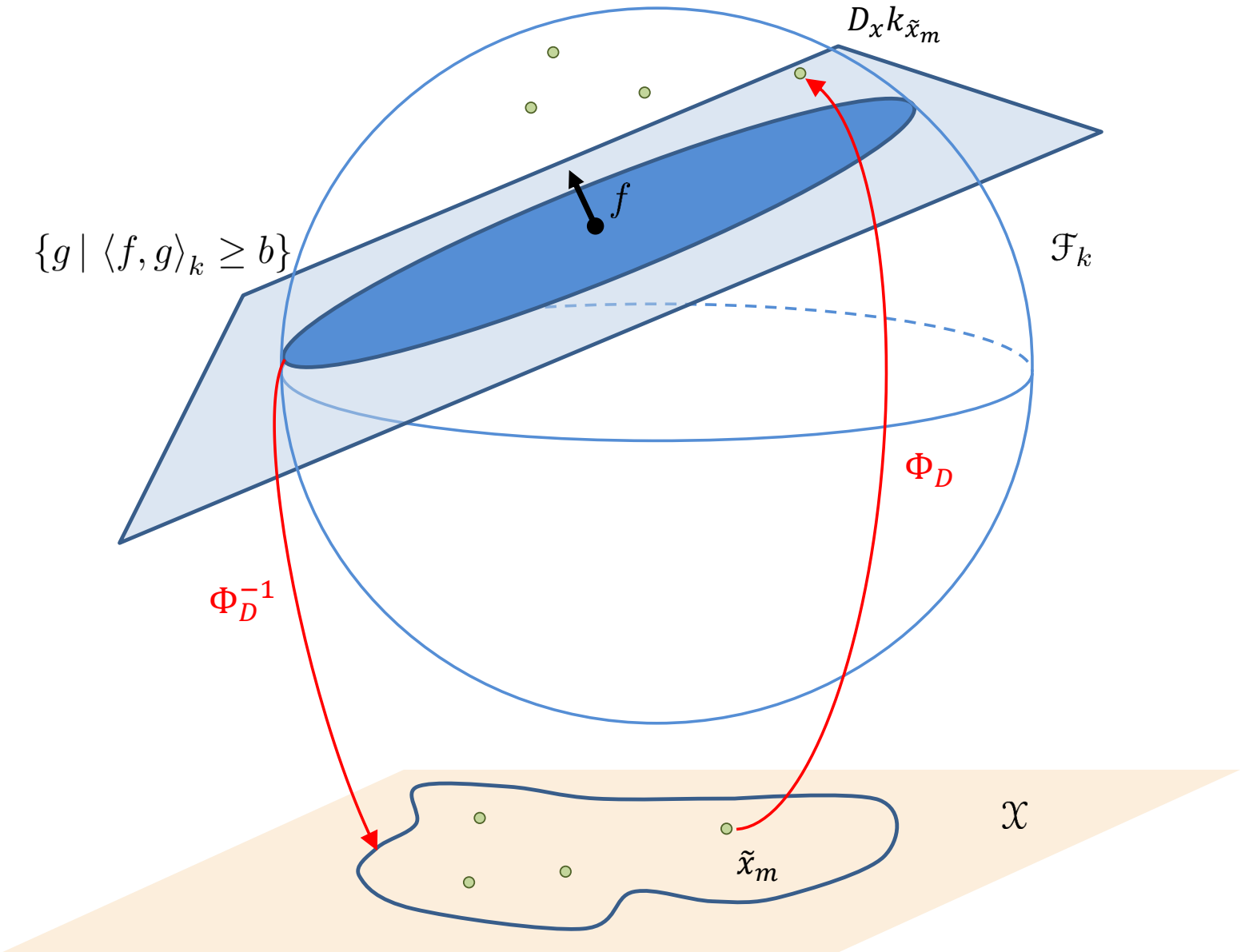
Shift-invariant kernel ($k(x, y) = k_0(x - y)$) gives

$$
\eta(\delta) = \sup_{u \in \mathbb{B}_{\|\cdot\|_{\mathcal{X}}}(0, \delta)} \sqrt{|2 D_x D_y k_0(0) - 2 D_x D_y k_0(u)|}
$$

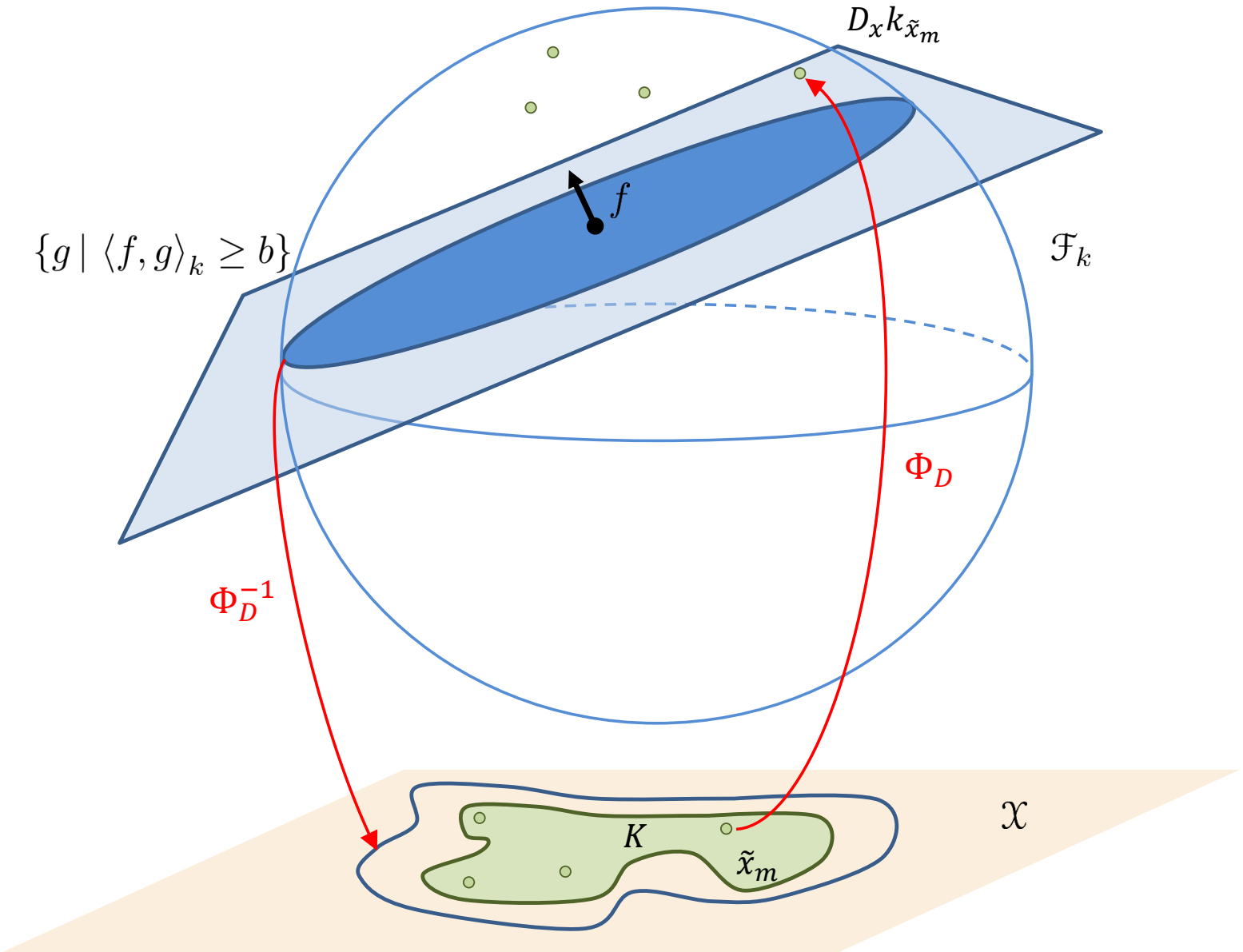
Other buffers are possible (e.g. constant), why choose "$\eta_m \|f(\cdot)\|$"?
$\hookrightarrow$ This choice comes from a geometrical interpretation.



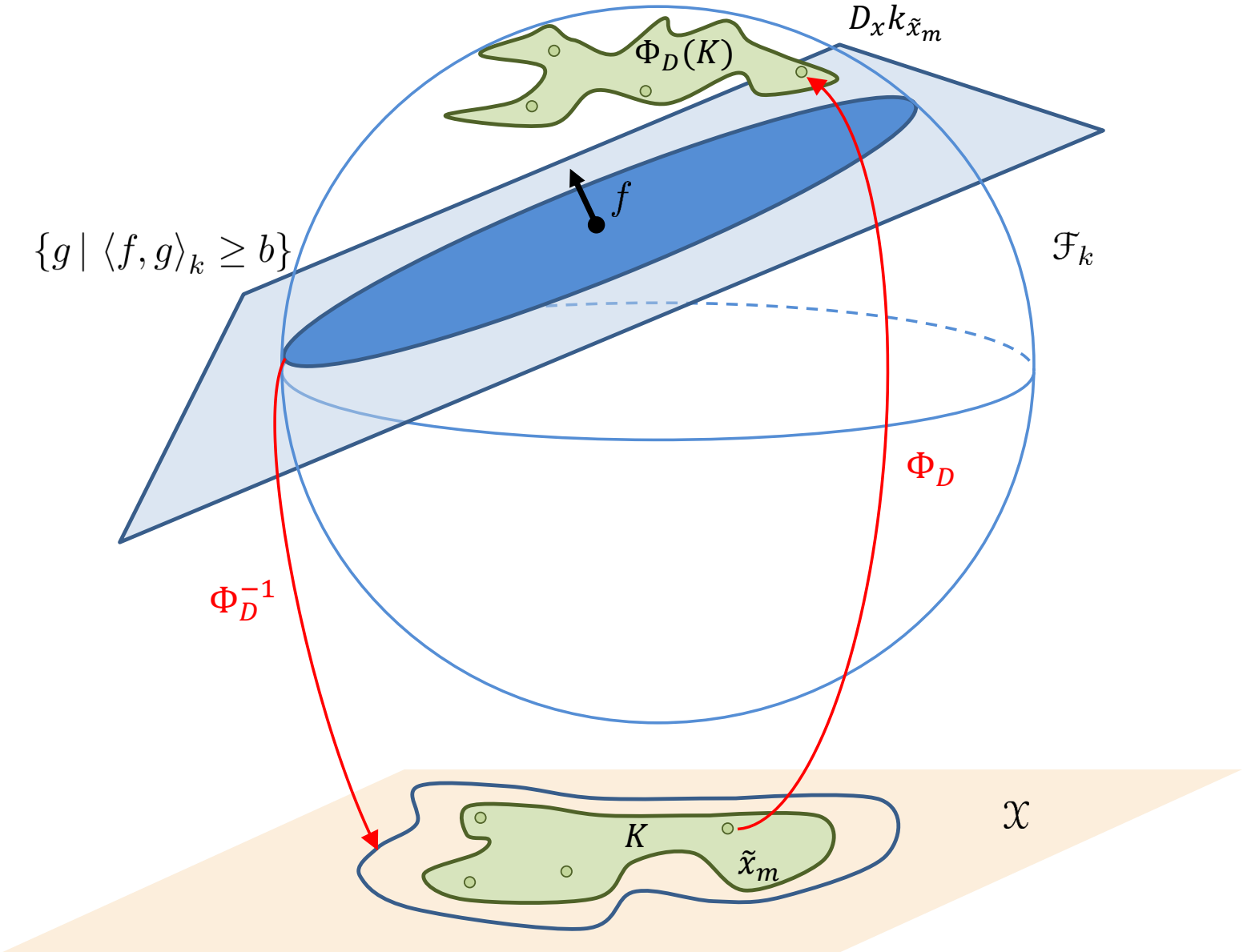
Support Vector Machine (SVM) is about separating red and green points by blue hyperplane.



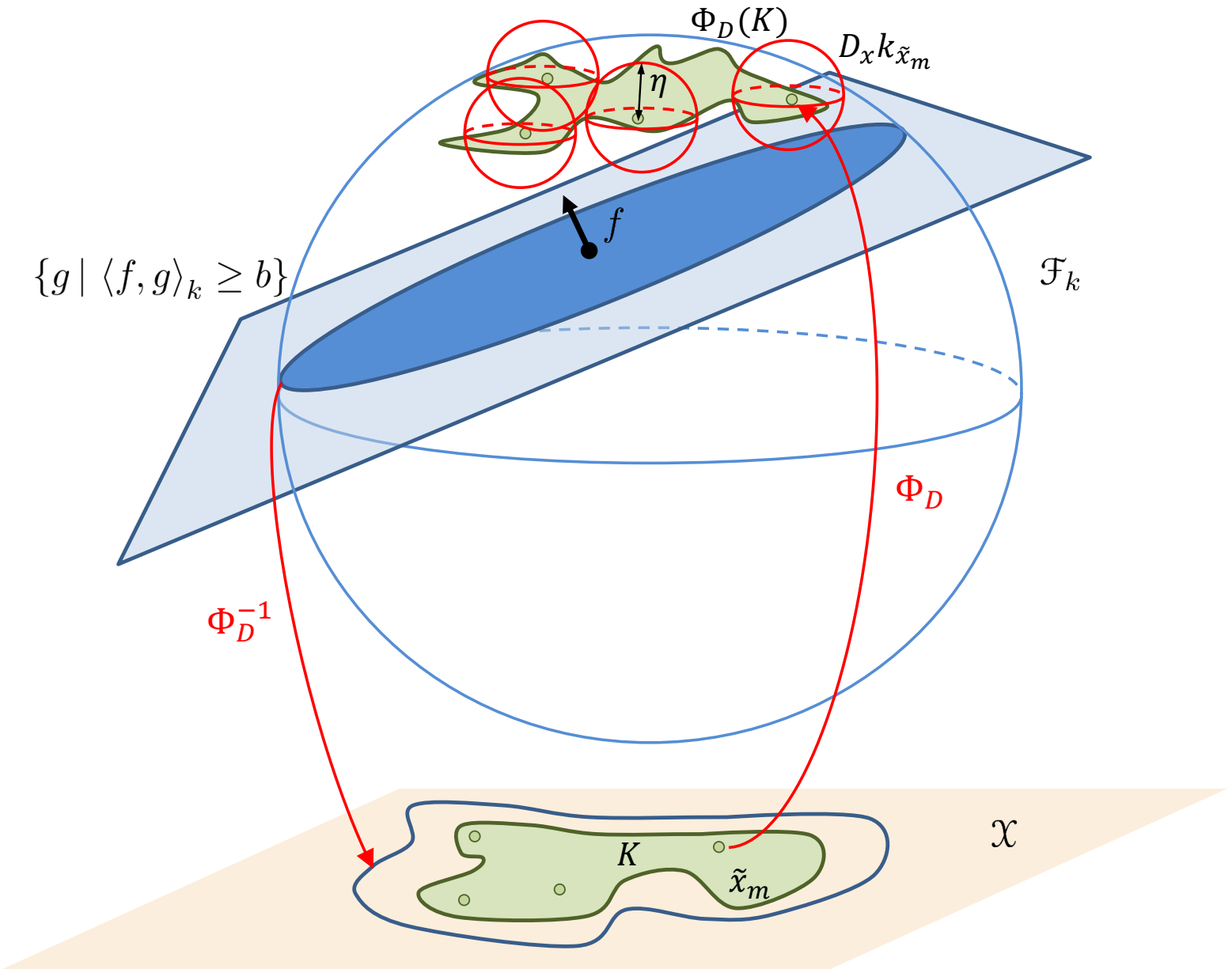
Using the nonlinear embedding $\Phi_D : x \mapsto D_x k(x, \cdot)$, the idea is the same. Consider only the green points, it looks like one-class SVM.



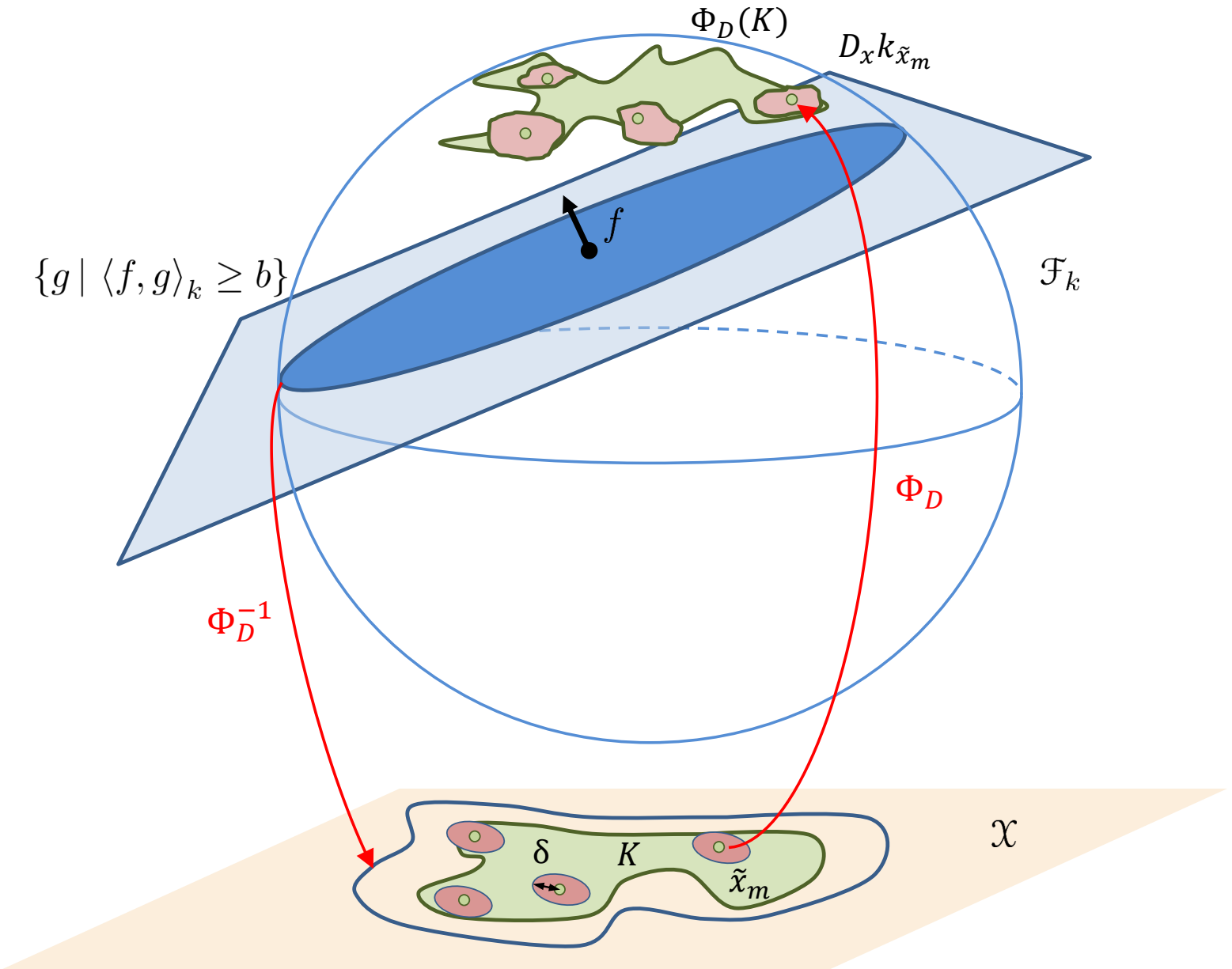
The green points are now samples of a compact set $\mathcal{K}$.



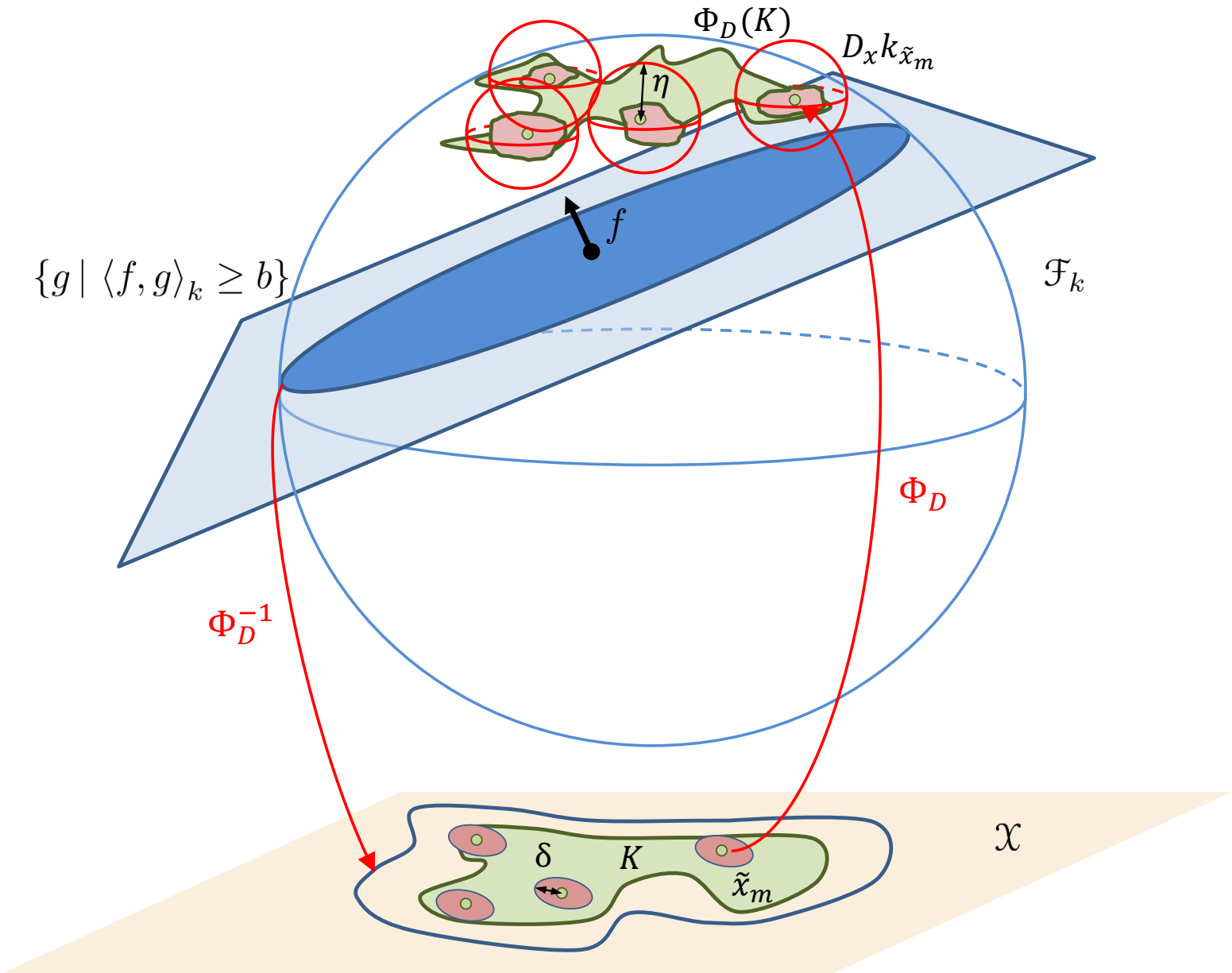
The image $\Phi_D(\mathcal{K})$ looks ugly...



The image $\Phi_D(\mathcal{K})$ looks ugly, can we cover it by balls? How to choose η?



First cover $\mathcal{K} \subset \bigcup\{\tilde{x}_m + \delta\mathbb{B}\}$, and then look at the images $\Phi_D(\{\tilde{x}_m + \delta\mathbb{B}\})$



Cover the $\Phi_D(\{\tilde{x}_m + \delta\mathbb{B}\})$ with tiny balls! This is how SOC was defined.

Main theorem

$$f_\eta \in \underset{f \in \mathcal{F}_k}{\arg\min}\ \mathcal{L}(f) = L\left((x_n, y_n, f(x_n))_{n\in[N]}\right) + \Omega\left(\|f\|_k\right)$$
$$\text{s.t.} \qquad b_i + \eta_{i,m}\|f(\cdot)\|_k \leq D_i f(\tilde{x}_{m,i}), \quad \forall\, m \in [M_i],\ \forall i \in [\mathcal{I}].$$

if $\Omega(\cdot)$ is strictly increasing, then

Theoretical guarantees [Aubin-Frankowski and Szabó, 2020]

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

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

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

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

Application: Joint Quantile Regression (JQR)

$f_\tau(x)$ conditional quantile over (X, Y): $P(Y \leq f_\tau(x)|X = x) = \tau \in]0, 1[$.
Estimation through convex optimization over "pinball loss" $l_\tau(\cdot)$ (i.e. tilted absolute value [Koenker, 2005]).

Joint quantile regression with non-crossing constraints, over $(f_q)_{q\in[Q]}$:

$$\mathcal{L}(f_1, \dots f_Q) = \frac{1}{N} \sum_{q\in[Q]} \sum_{n\in[N]} l_{\tau_q}(y_n - f_q(x_n)) + \lambda_f \sum_{q\in[Q]} \|f_q\|_k^2$$

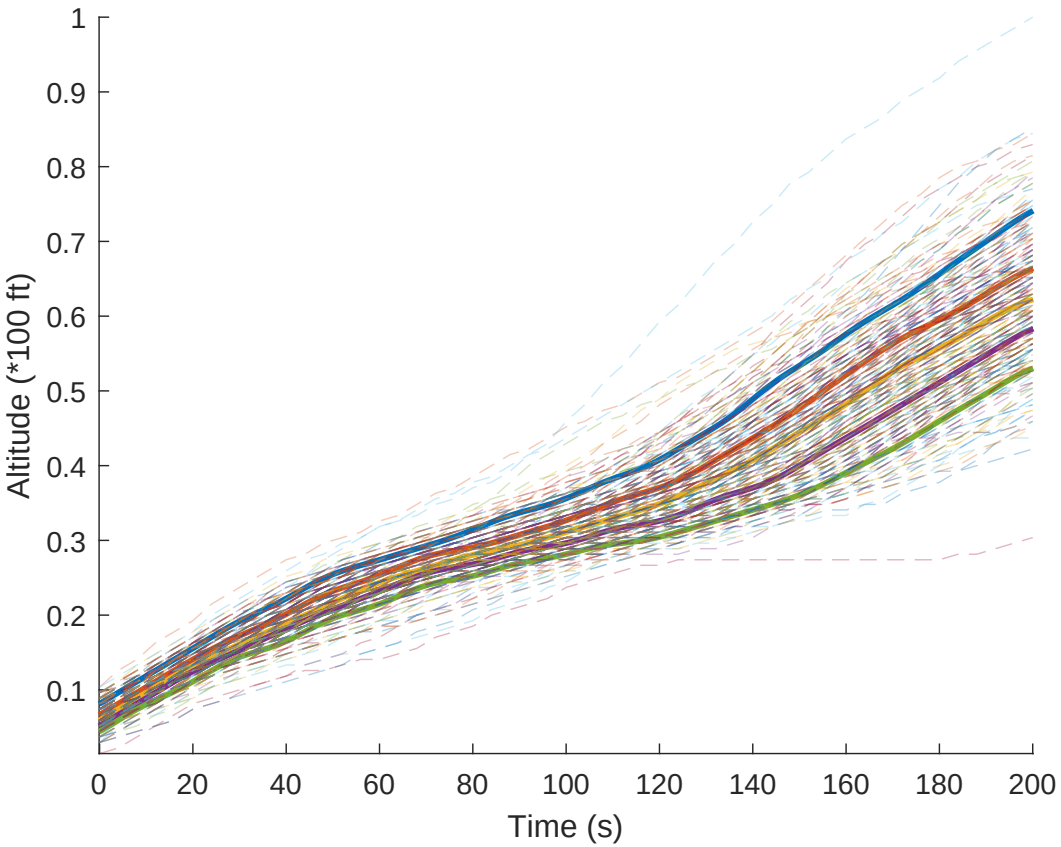
$$\text{s.t. } f_{q+1}(x) \geq f_q(x), \forall q \in [Q-1], \forall\, x \in [\min x_n, \max x_n]^d.$$

Known fact: quantile functions can cross when estimated independently.

Can we pair the non-crossing constraint with other physical requirements?

Joint quantile regression (JQR): airplane data

Airplane trajectories at takeoff have increasing altitude.



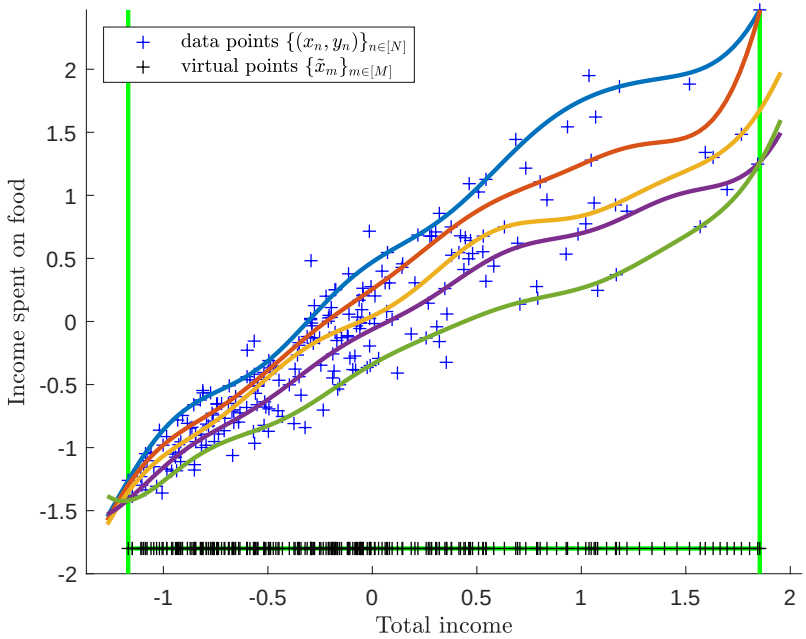
JQR with monotonic constraint over $[x_{\min}, x_{\max}]$:

Increasing quantiles should be non-crossing

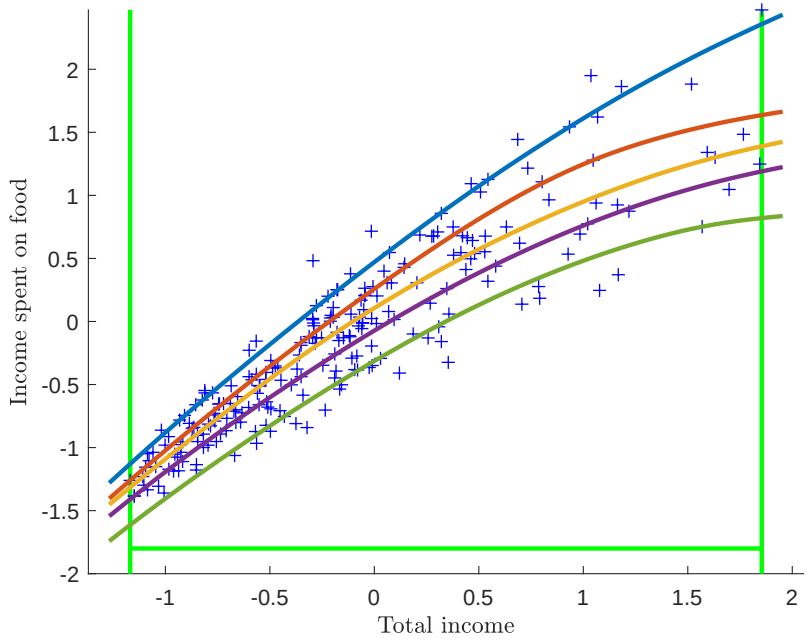
Two shape constraints jointly handled with $15k$ samples.
Works with higher dimensions too!

Joint quantile regression (JQR): Engel's law

Engel's law (1857): "As income rises, the proportion of income spent on food falls, but absolute expenditure on food rises."



Increasing+non-crossing

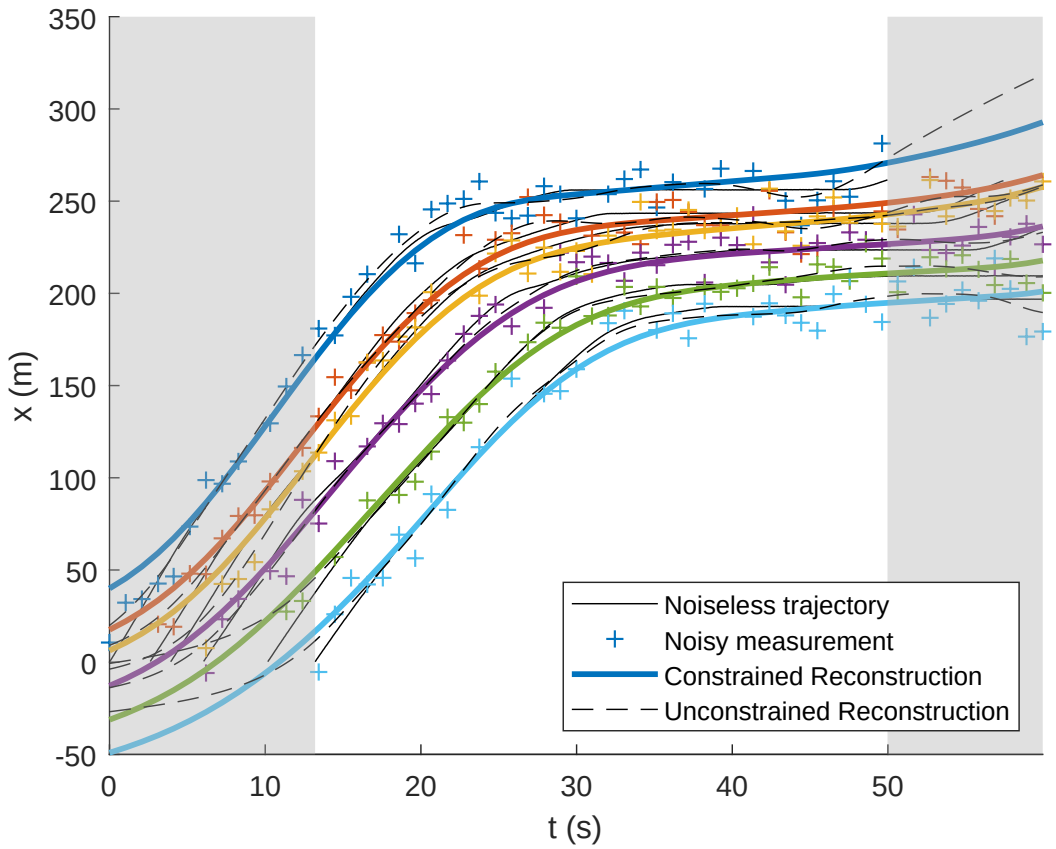


Increasing+non-crossing+concave

Priors have a great effect on the shape of solutions!

Kernel ridge regression (KRR): trajectory reconstruction

Very noisy GPS data: six non-overtaking cars in a traffic jam



KRR with monotonic constraint over $[t_{\min}, t_{\max}]$:

Forward trajectories also maintain security distance

(In *Kernel Regression for Vehicle Trajectory Reconstruction under Speed and Inter-vehicular Distance Constraints*, PCAF and Nicolas Petit and Zoltán Szabó *IFAC World Congress 2020*)

Teaser slide

This approach works as well for

- Other compact coverings than balls
- SDP constraints (e.g. convexity for $d \geq 2$): $0 \preccurlyeq \mathbf{Hess}(f)(x)$
- Vector-valued functions $f : \mathcal{X} \to \mathbb{R}^Q$
- Other applications: finance, control theory, ...

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

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

Conclusion/Take-home message

- tightening intractable constraints is the only way to have guarantees
- compact coverings in infinite dimensional spaces can be useful

See *Hard Shape-Constrained Kernel Machines*, PCAF and Zoltán Szabó, June 2020, `https://arxiv.org/abs/2005.12636`

Acknowledgments

Thank you for your attention!

Deeply grateful to:

Nicolas Petit (Mines ParisTech), my PhD advisor

Jean-Philippe Vert (Mines ParisTech-Google), who got me to love kernels

Zoltán Szabó[1] (Ecole polytechnique), my co-author for quantile regression

INSTITUT Louis Bachelier

Stress Test
RISK Management and Financial Steering

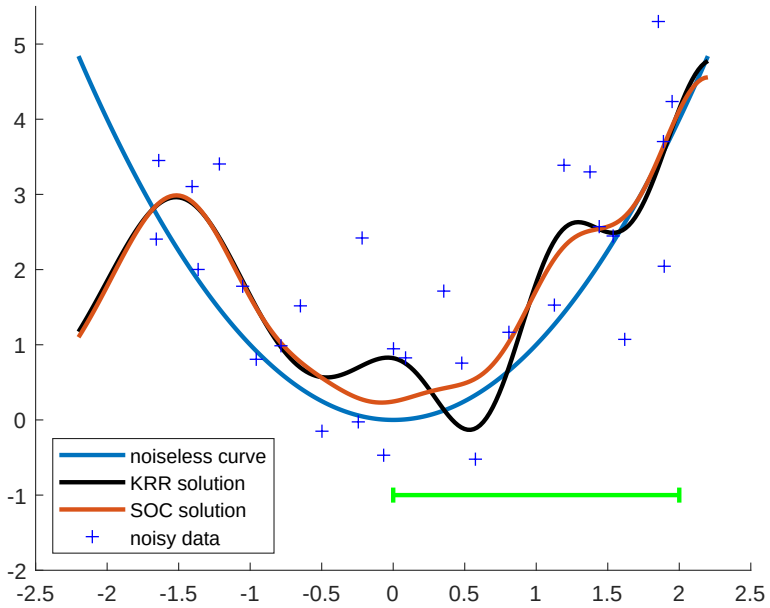
[1] ZSz benefited from the support of the Europlace Institute of Finance and that of the Chair Stress Test, RISK Management and Financial Steering, led by the French École Polytechnique and its Foundation and sponsored by BNP Paribas.

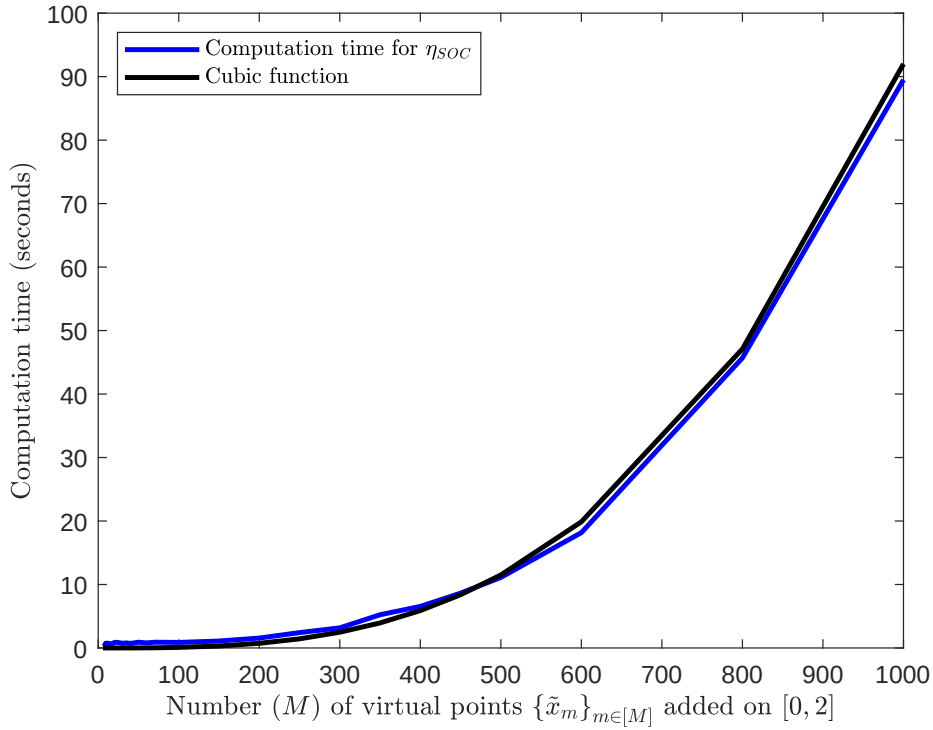
Appendix: JQR performance over UCI datasets

- PDCD = Primal-Dual Coordinate Descent [Sangnier et al., 2016], JQR with parallel/heteroscedatic quantile penalization (see also ITL [Brault et al., 2019] for noncrossing inducer)
- mean $\pm$ std of 100$\times$value of the pinball loss (smaller is better)

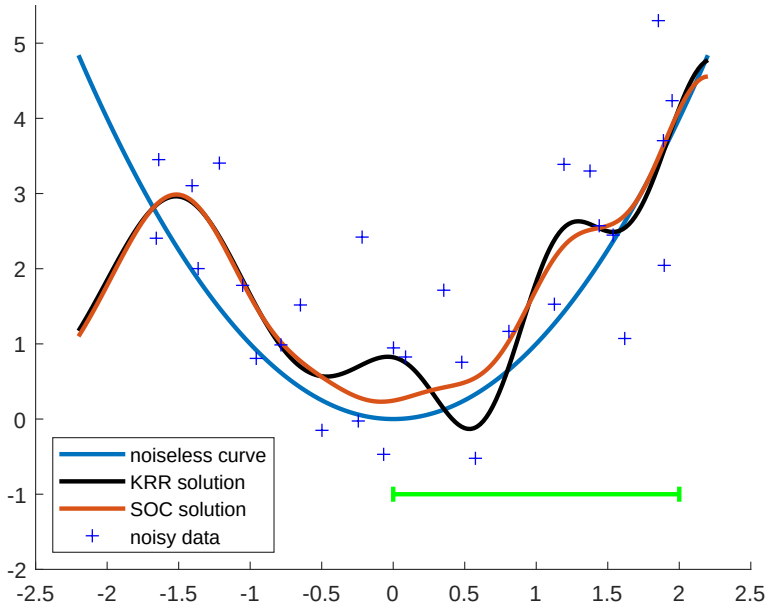
Dataset	d	N	PDCD	SOC
engel	1	235	48 ± 8	53 ± 9
GAGurine	1	314	61 ± 7	65 ± 6
geyser	1	299	105 ± 7	108 ± 3
mcycle	1	133	66 ± 9	62 ± 5
ftcollinssnow	1	93	154 ± 16	148 ± 13
CobarOre	2	38	159 ± 24	151 ± 17
topo	2	52	69 ± 18	62 ± 14
caution	2	100	88 ± 17	98 ± 22
ufc	3	372	81 ± 4	87 ± 6

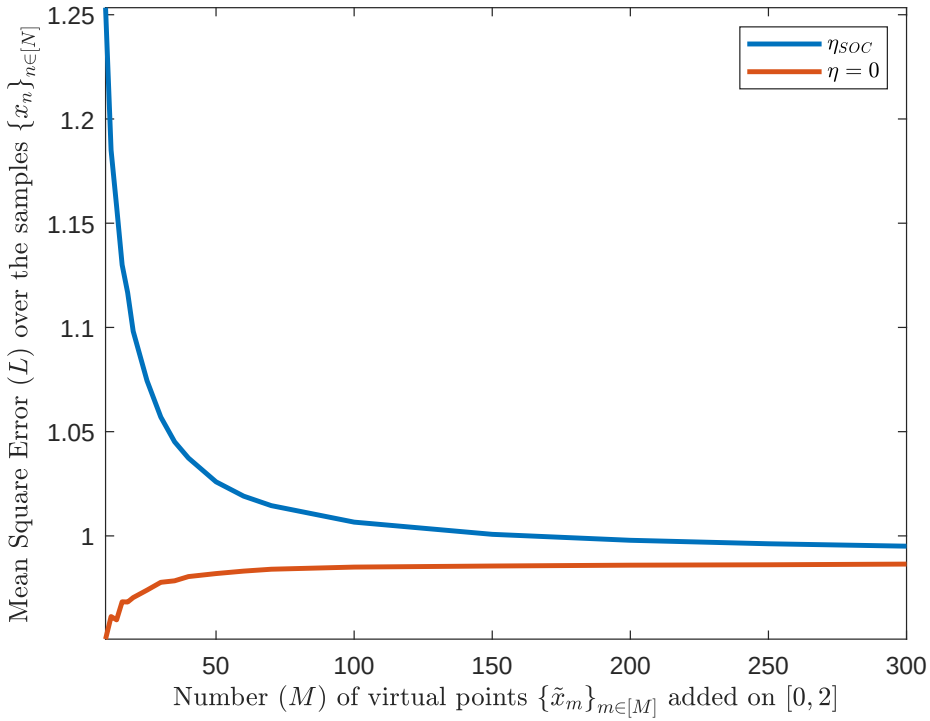
Appendix: KRR time computation with increasing number of virtual points M





Appendix: KRR performance with increasing number of virtual points M





References I

Agrell, C. (2019).

Gaussian processes with linear operator inequality constraints.

Journal of Machine Learning Research, 20:1–36.

Aronszajn, N. (1950).

Theory of reproducing kernels.

Transactions of the American Mathematical Society, 68:337–404.

Aubin-Frankowski, P.-C. and Szabó, Z. (2020).

Hard shape-constrained kernel machines.

https://arxiv.org/abs/2005.12636.

Brault, R., Lambert, A., Szabo, Z., Sangnier, M., and d'Alche Buc, F. (2019).

Infinite task learning in RKHSs.

volume 89 of *Proceedings of Machine Learning Research*, pages 1294–1302. PMLR.

References II

Egerstedt, M. (2009).

Control Theoretic Splines: Optimal Control, Statistics, and Path Planning (Princeton Series in Applied Mathematics).

Princeton University Press.

Hall, G. (2018).

Optimization over nonnegative and convex polynomials with and without semidefinite programming.

PhD Thesis, Princeton University.

Koenker, R. (2005).

Quantile Regression.

Econometric Society Monographs. Cambridge University Press.

Matzkin, R. L. (1991).

Semiparametric estimation of monotone and concave utility functions for polychotomous choice models.

Econometrica, 59(5):1315–1327.

References III

Papp, D. and Alizadeh, F. (2014).

Shape-constrained estimation using nonnegative splines.

Journal of Computational and Graphical Statistics, 23(1):211–231.

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.

Simchi-Levi, D., Chen, X., and Bramel, J. (2014).

The Logic of Logistics: Theory, Algorithms, and Applications for Logistics Management.

Springer.

Takeuchi, I., Le, Q., Sears, T., and Smola, A. (2006).

Nonparametric quantile estimation.

Journal of Machine Learning Research, 7:1231–1264.