Functional Optimization by Variable-Basis Approximation Schemes

Giorgio Gnecco

DIMA (Department of Mathematics)
\&
DIST (Department of Communication, Computer and System Science)

University of Genoa

Ph.D. Advisor: Marcello Sanguineti
Part I

Overview
Functional optimization problems

Problem formulation:

- $(\mathcal{X}, \| \cdot \|_\mathcal{X})$: normed linear space of functions (ambient space).
- $\Phi : \mathcal{X} \to \mathbb{R}$: functional.
- $S \subseteq \mathcal{X}$: set of admissible functions.

**Problem $(S, \Phi)$:**

$$\inf_{f \in S} \Phi(f).$$

(related problem: properties of a minimizer $f^o$ (if the inf is achieved))

Often:

- **Infinite-dimensional** ambient space.
- **Nonlinear** functional.
- $S$ is a family of real-valued functions defined on a set $X \subseteq \mathbb{R}^d$, with $d \gg 1$ (large number of variables).
Examples of functional optimization problems

For suitable optimality criteria:

- Find optimal exploration strategies in a partially-unknown environment.
- Find optimal routing strategies in telecommunication networks.
- Optimize a consumer’s choice under uncertainty.
- Optimize the strategies of multiple decisors with a common objective but different available information (team decision problems).
- Find an optimal smooth interpolation or approximation of a data set \((x_i, y_i), \ i = 1, \ldots, m\) (typically, \(x_i \in X \subseteq \mathbb{R}^d\) and \(y_i \in \mathbb{R}\)).
Fixed-basis and variable-basis approximation schemes (I)

- \( \{S_n\} \) nested sequence of subsets of \( S \)
- \( n \in \mathbb{N}^* \) controls the "level of sparsity". E.g.,
- \( \mathcal{X}_1 \subset \mathcal{X}_2 \subset \ldots \subset \mathcal{X}_n \subset \ldots \subset \mathcal{X} \), \( n \)-dimensional subspaces, and \( S_n = S \cap \mathcal{X}_n \) (fixed-basis approximation scheme).
- \( G \subset \mathcal{X} \) given and \( S_n = S \cap \text{span}_n G \) (variable-basis approximation scheme or vbs), where

\[
\text{span}_n G := \{ f \in \mathcal{X} : f = \sum_{i=1}^{n} c_i g_i, \ c_i \in \mathbb{R}, \ g_i \in G \}.
\]

- Vbs model neural networks: e.g., NN with sigmoidal or Gaussian computational units.
- \( \inf_{f \in S \cap \text{span}_n G} \Phi(f) \): Extended Ritz Method (ERIM).
Quantities of interest (assume that an optimal $f^o$ exists and is unique):

$$\inf_{f \in S_n} \Phi(f) - \inf_{f \in S} \Phi(f),$$

$$\sup_{f \in \text{argmin}_\varepsilon(S_n, \Phi)} \|f - f^o\|_\mathcal{X}$$

(for $\varepsilon > 0$, $\text{argmin}_\varepsilon(S_n, \Phi) := \{f \in S_n : \Phi(f) \leq \inf_{g \in S_n} \Phi(g) + \varepsilon\}$).

Good choice of $\{S_n\}$: guided by structural properties of $f^o$. 
Rates of approximate optimization & the curse of dimensionality

- **Effective** sparse suboptimal solutions:

  \[
  \inf_{f \in S_n} \Phi(f) - \inf_{f \in S} \Phi(f) \leq Cr(n) \quad \text{(upper bound)},
  \]

  where \( C > 0 \), \( r : \mathbb{N}^* \to \mathbb{R}^+ \) and \( \lim_{n \to +\infty} r(n) = 0 \) “sufficiently fast”.

- **Unfortunately**, for some families of problems:

  \[
  C_1(d) \left( \frac{1}{n} \right)^{C_2(d)} \leq \inf_{f \in S_n} \Phi(f) - \inf_{f \in S} \Phi(f) \quad \text{(lower bound)},
  \]

  where \( S \) is a class of real-valued functions on \( X \subset \mathbb{R}^d \), \( C_1 : \mathbb{N}^* \to \mathbb{R}^+ \), and \( C_2 : \mathbb{N}^* \to \mathbb{R}^+ \) is a linear function in \( d \).

  >>> **Curse of dimensionality** (Bellman, 1957) in the size of the approximation scheme (or, shortly, in \( n \)).
Why variable-basis functions?

- **Theoretical motivations.** For certain function approximation problems in real Hilbert spaces \((\mathcal{H}, \| \cdot \|_\mathcal{H})\), for which
  - \(S = \mathcal{H}\),
  - \(\bar{f} : X \subseteq \mathbb{R}^d\) is a function belonging to a suitable subset of \(\mathcal{H}\),
  - \(\Phi(f) = \| f - \bar{f} \|_\mathcal{H}\),
  - \(\inf_{f \in S} \Phi(f) = 0\),

(Barron, 1993) showed that the best fixed-basis approximation schemes suffer from the curse of dimensionality in \(n\), while for certain vbs

\[
\inf_{f \in \text{span}_n G} \| f - \bar{f} \|_\mathcal{H} \leq C(d) \sqrt{\frac{1}{n}},
\]

where (at least) the second factor does not depend on \(d\).

Classes of problems for which \(C(d)\) grows “slowly” with \(d\) (Kainen, Kůrková, & Sanguineti, 2008).

- **Experimental motivations** for other functional optimization problems (ERIM).
Main objectives of the thesis

- Deriving structural properties of the unknown optimal solutions to functional optimization problems.

- Exploiting such structural properties to choose families of variable-basis approximators, which
  - are sparse;
  - contain sufficiently accurate suboptimal solutions;
  - require a number $n$ of basis functions that does not grow “too fast” with the number of variables of the admissible solutions.

We are interested in

- studying families of problems for which the curse of dimensionality in the size of the approximation scheme is avoided or at least mitigated.
The curse of dimensionality can show up also in the minimum

- memory,
- time,
- and available information (e.g., size of the data sample),

required to achieve a desired degree of accuracy of suboptimal solutions.

→ apply tools from Information-based Complexity.
Examples of structural properties

For $X \subseteq \mathbb{R}^d$, $k \in \mathbb{N}^*$, and $p \in [1, +\infty]$:
- $f^o \in \mathcal{C}^k(X)$, $f^o \in \mathcal{W}^{k,p}(X)$, or $f^o \in \mathcal{B}^{k,p}(X)$ (Bessel potential space).

(strong) convexity of $f^o$ (e.g., when $f^o$ is the value function in certain (strong) convex dynamic optimization problems).

$f^o$ is in the closure of the symmetric convex hull of a set of functions $F$:

$$f^o \in \text{cl conv}(F \cup -F).$$
Original contributions: summary(I)

- New approximation bounds for vbs (Chapter 2), exploited in subsequent chapters for certain functional optimization problems (via the moduli of continuity of the associated functionals).
- Estimates of $G$-variation norm (Chapter 7).
- For deterministic $N$-stage optimization problems (Chapters 3 and 4):
  - smoothness properties of optimal solutions (Chapter 3);
  - accuracy of sparse suboptimal solutions by the ERIM (vbs approximation of the $t$-th stage optimal policies) (Chapter 3);
  - accuracy of sparse suboptimal solutions by approximate dynamic programming (vbs approximation of the $t$-th stage value functions) (Chapter 4).
For stochastic team optimization problems (Chapter 5):
- smoothness properties of optimal solutions;
- accuracy of sparse suboptimal solutions by the ERIM (vbs approximation of the optimal strategies).

Learning from data (Chapter 6):
- accuracy of sparse suboptimal solutions by the ERIM, for various regularization techniques (e.g., Tikhonov, Ivanov, Miller, and Phillips regularizations, weight decay).

Not included in the thesis:
- smoothness properties for stochastic $N$-stage optimization problems;
- accuracy of suboptimal solutions by vbs, for Kernel Principal Component Analysis (KPCA);
- Statistical Learning Theory bounds for sparse suboptimal solutions in learning from data.
Functional Optimization by Variable-Basis Approximation Schemes

Overview

Novelties with respect to previous works on vbs

Novelties with respect to previous works on vbs (I)

A large part of the thesis is devoted to obtain upper bounds on

$$\inf_{f \in S \cap \text{span}_n G} \Phi(f) - \inf_{f \in S} \Phi(f) :$$

- theoretical study;
- applications in Operations Research.

We investigate dynamic optimization problems, instead of static ones.

We investigate also optimization problems with multiple decision makers and a common objective.
Novelties with respect to previous works on vbs (II)

- **Structural properties:**
  - Previous works on optimization by vbs assume that an optimal solution $f^*$ satisfies them (e.g., $f^0 \in \text{cl conv}(G \cup -G)$).
  - Instead, we investigate conditions on $S$ and $\Phi$, guaranteeing certain structural properties.
  - Factorized estimates

$$\inf_{f \in S \cap \text{span}_n G} \Phi(f) - \inf_{f \in S} \Phi(f) \leq C(d) \sqrt{\frac{1}{n}}.$$ 

for sufficiently large degrees of smoothness in the problems’ formulations.

- The so-called “blessing of smoothness” (a term coined by Girosi (Poggio and Smale, 2004)) helps to mitigate the curse of dimensionality in the size of the approximation scheme.
Novelties with respect to previous works on vbs (III)

- **Function approximation by vbs:**
  - Previous works exploit upper bounds on the modulus of continuity of the functional $\Phi$ at $f^*$ (to apply, to the functional optimization problem $(S, \Phi)$, an upper bound for scalar-valued function approximation (Maurey-Jones-Barron’s lemma)).

- **We derive new upper bounds:**
  - for scalar function approximation by vbs;
  - specific to vector function approximation by vbs (for the approximation of the optimal policies in dynamic optimization problems).
Novelties with respect to previous works on vbs (IV)

- **New estimates of \(G\)-variation norm** (a norm associated with vbs):
  - Previous works restate MJB’s lemma in terms of a norm called \(G\)-variation (Kůrková, 1997), associated with the set of functions \(G\) used in the vbs.
  - When a function \(f\) has an integral representation depending on a kernel function and a weight function \(w \in L^1\), (Kůrková et al., 1997) proved that the \(G\)-variation of \(f\) is bounded from above by the \(L^1\)-norm of \(w\).
  - We investigate conditions under which this upper bound is an equality.
  - We derive upper bounds on the \(G\)-variation of \(f\), in terms of “classical” norms.
International journals

1. G. Gnecco and M. Sanguineti.
   Suboptimal solutions to dynamic optimization problems via approximations of the policy functions.

2. G. Gnecco and M. Sanguineti.
   Estimates of variation with respect to a set and applications to functional optimization problems.

   The weight-decay technique in learning from data: An optimization point of view.

   Accuracy of suboptimal solutions to kernel principal component analysis.

5. A. Alessandri, G. Gnecco, and M. Sanguineti.
   Computationally Efficient Approximation Schemes for Functional Optimization.
Value and policy function approximations in infinite-horizon optimization problems.


Approximation error bounds via Rademacher's complexity.

Book chapters

1. G. Gnecco and M. Sanguineti.
Regularization issues and suboptimal solutions in learning from data.

2. A. Alessandri, G. Gnecco, and M. Sanguineti.
Computationally Efficient Approximation Schemes for Functional Optimization.
Publications (III)

International conferences

1. G. Gnecco and M. Sanguineti.
   Suboptimal solutions to network team optimization problems.

   Exploiting Structural Results in Approximate Dynamic Programming.
   *22nd European Conf. on Operational Research (EURO 2007)*.

National conferences

1. G. Gnecco and M. Sanguineti.
   Structural Properties of Stochastic Dynamic Concave Optimization Problems and Approximations of the Value and Optimal Policy Functions.
   *XXXIX Annual Conf. of the Italian Operations Research Society (AIRO), 2008*.

   *XXXVIII Annual Conf. of the Italian Operations Research Society (AIRO), 2007*.
*Deriving Approximation Error Bounds via Rademacher Complexity and Learning Theory.*

*Accuracy of Suboptimal Solutions to Kernel Principal Component Analysis.*
Part II

Selection of the results
Approximation error bounds for vbs (I)

- **Hint**: Girosi’s idea to exploit the Vapnik-Chervonenkis Theorem (Girosi, 1995).

- **We improve them** by using more recent bounds from Statistical Learning Theory, based on the concept of Rademacher’s complexity of a family of (scalar) real-valued functions.

- We bound the Rademacher’s complexity from above in terms of the Vapnik-Chervonenkis dimension.

- **We extend the results** to families of vector-valued functions, exploring the possibility of reducing the number of parameters to be optimized, when there are similarities among their components.
Rademacher’s complexity

- **Rademacher’s random variable**: random variable taking only the values $+1$ and $-1$ with equal probability $1/2$.
- $\{\varepsilon_i\}, \ i = 1, \ldots, n$: sequence of $n$ independent Rademacher’s random variables.
- $\{t_i\}, \ i = 1, \ldots, n$: sequence of $n$ independent identically distributed samples, drawn from a probability measure $P_X$ on a set $X$.
- $F$: family of real-valued functions defined on $X$.

Rademacher’s complexity of $F$:

$$R_n(F) := \mathbb{E}_{t_1, \ldots, t_n} \mathbb{E}_{\varepsilon_1, \ldots, \varepsilon_n} \left\{ \frac{1}{\sqrt{n}} \sup_{f \in F} \left| \sum_{i=1}^{n} \varepsilon_i f(t_i) \right| \right\}.$$
VC dimension for a family of real-valued functions

- Vapnik-Chervonenkis dimension (VC dimension) of a family $F$ of real-valued functions on $X$:
  - maximum cardinality of a set $\{t_i, i = 1, \ldots, h\} \subseteq X$ that can be partitioned into two classes $\Omega_1$ and $\Omega_2$ in all $2^h$ possible ways, by using functions of the form $f - \alpha$, with $f \in F$ and $\alpha \in \mathbb{R}$.
  - If $f(t_i) - \alpha \geq 0$, then $(f, \alpha)$ assigns $t_i$ to the class $\Omega_1$.
  - If $f(t_i) - \alpha < 0$, then $(f, \alpha)$ assigns $t_i$ to the class $\Omega_2$.

Lemma 2.3.4

Let $F$ be a family of $[0, 1]$-valued functions with finite VC dimension $\text{VC}(F)$. Then there exists an absolute positive constant $C$ such that for every positive integer $n$

$$\mathcal{R}_n(F) \leq C \sqrt{\text{VC}(F)}.$$
**Theorem 2.3.5**

Let $X \subseteq \mathbb{R}^d$ be nonempty, $\lambda \in L^1(X)$, $K : X \times X \to \mathbb{R}$ a kernel such that there exists $\tau > 0$ with $|K(x, t)| \leq \tau$ for every $x, t \in X$, $f : X \to \mathbb{R}$ a function with the integral representation $f(x) = \int_X K(x, t) \lambda(t) \, dt$, and $h$ the VC dimension of the family $\{K(x, \cdot)\}$. Then there exists an absolute positive constant $C$ such that, for every positive integer $n$, there exist $t_1, \ldots, t_n \in X$ and $c_1, \ldots, c_n \in \{-1, 1\}$ such that

$$
\sup_{x \in X} \left| f(x) - \frac{\|\lambda\|_1}{n} \sum_{i=1}^n c_i K(x, t_i) \right| \leq C \|\lambda\|_1 \sqrt{\frac{h}{n}}.
$$
Approximation error bounds for vbs (III)

\[ C \text{ may be difficult to estimate, however the upper bound is asymptotically better than the one obtained by Girosi:} \]

\[
\sup_{x \in X} \left| f(x) - \frac{\|\lambda\|_1}{n} \sum_{i=1}^{n} c_i K(x, t_i) \right| \leq 4\tau \|\lambda\|_1 \sqrt{\frac{h \ln \frac{2en}{h} + \ln 4}{n}}.
\]

(in general, \( c_i \) and \( t_i \) different from those in previous theorem).

\[
\sqrt{\frac{h}{n}} \quad \text{versus} \quad \sqrt{\frac{h \ln \frac{2en}{h} + \ln 4}{n}}.
\]
Extension to vector-valued functions

- $X \subset \mathbb{R}^d$ compact and $f : X \rightarrow \mathbb{R}^k$ such that

$$f_m(x) = \int_X K_m(x, t)\lambda(t)dt, \quad m = 1, \ldots, k,$$

where $\lambda(t) \in \mathcal{L}^1(X)$, $K_m(x, t) = K(x, t)\lambda_m(t)$ for $K$ continuous with $\max_{(x, t) \in X \times X} |K(x, t)| \leq \tau$ and each $\|\lambda_m\|_\infty \leq 1$.

**Theorem 2.5.4**

*For every $n \in \mathbb{N}^*$, there exist $t_1, \ldots, t_n \in X$ and $c_{m,1}, \ldots, c_{m,n} \in \{-1, 1\}$ $(m = 1, \ldots, k)$ such that*

$$\sup_{x \in X} \left| f_m(x) - \frac{\|\lambda\|_1}{n} \sum_{i=1}^n c_{m,i} K(x, t_i) \right| \leq 2 \tau C \|\lambda\|_1 \sqrt{\frac{1}{n} \max \left\{ \left( \frac{\mathcal{R}_n + \tau}{2\tau} \right)^2, \ln k \right\}}.$$

Note that the $t_i$ are the same for each component of $f$. 
Dynamic $N$-stage deterministic optimization problems (I)

- Problem $\Sigma_N$: For each $x_0 \in X \subseteq \mathbb{R}^d$, find

$$J^*(x_0) = \sup \left\{ \sum_{t=0}^{N-1} \beta^t h(x_t, x_{t+1}) + \beta^N h_N(x_N) \right\}$$

$$s.t. \quad (x_t, x_{t+1}) \in D, \quad t = 0, 1, \ldots, N-1, \quad x_t \in X .$$

- $N$ is a given finite horizon.
- $D \subseteq X \times X$.
- $\beta > 0$ is a fixed discount factor.
- $h : D \rightarrow \mathbb{R}$, $h_N : X \rightarrow \mathbb{R}$: transition reward and final reward.
Formulation in terms of optimal policies

- For every \( x \in X \): \( D(x) := \{ y \in X : (x, y) \in D \} \).
- **Policy function**: \( g : X \rightarrow X \) such that \( g(x) \in D(x) \) for each \( x \in X \).
- Under mild conditions, Problem \( \Sigma_N \) is equivalent to
  
  Problem \( \Sigma'_N \): Find policy functions \( g_0, \ldots, g_{N-1} \) such that
  
  \[
  J^o(x_0) = \sup \left\{ \sum_{t=0}^{N-1} \beta^t h(x_t, x_{t+1}) + \beta^N h_N(x_N) \right\}
  \]
  
  s.t. \( x_{t+1} = g_t(x_t), \ t = 0, 1, \ldots, N - 1, \ x_t \in X \).

→ We prove smoothness properties of the \( t \)-th optimal policy functions;
→ We exploit them to derive bounds for their approximation by Gaussian vbs.
\( \delta \)-concavity

- \( X \subseteq \mathbb{R}^d \) convex.
- \( \delta \in \mathbb{R} \).

\( f : X \rightarrow \mathbb{R} \) is \( \delta \)-concave on \( X \) if \( f(x) + \frac{1}{2} \delta \|x\|^2 \) is concave on \( X \).

- \( f \) of class \( C^2(X) \): necessary and sufficient condition for its \( \delta \)-concavity

\[
\sup_{x \in X} \lambda_{\text{max}}(\nabla^2 f(x)) \leq -\delta,
\]

where \( \lambda_{\text{max}}(\nabla^2 f(x)) \) is the maximum eigenvalue of the Hessian \( \nabla^2 f(x) \).
Assumption 3.4.1

Let $m$ be a positive integer. The following hold:
(i) $X \subset \mathbb{R}^d$ and $D \subseteq X \times X$ are compact, convex, and have nonempty interiors;
(ii) there exist optimal policies $g_{N-1}^o, \ldots, g_0^o$ that are continuous and interior on $X$, i.e., for every $x_t \in X$, $g_t^o(x_t) \in \text{int}(D(x_t))$, $t = N - 1, \ldots, 0$;
(iii) $h$ is concave and $h \in C^{m+1}(D)$, $h_N$ is $\delta$-concave for $\delta > 0$, and $h_N \in C^{m+1}(X)$;
(iv) there exists $\alpha > 0$ such that for every $a, b \in \mathbb{R}^d$ and every $(x, y) \in D$

\[ (a, b)' \nabla^2 h(x, y)(a, b) \leq -\alpha \|a\|^2. \]

See Section 3.4.2 for sufficient conditions for the continuity and interiority of $g_{N-1}^o, \ldots, g_0^o$. 
Optimal-policies approximation (II)

- $\mathcal{B}^{m,p}(\mathbb{R}^d)$: Bessel potential space, whose elements are functions $u : \mathbb{R}^d \to \mathbb{R}$ such that $u = f \ast G_m$, where $f \in L^p(\mathbb{R}^d)$ and $G_m$ is the Bessel potential of order $m$.

**Theorem 3.4.2**

Let Assumption 3.4.1 hold. If $1 < p < \infty$, then for every integer $t = 0, \ldots, N - 1$ and every $j = 1, \ldots, d$ there exists a function $\bar{g}_{t,j}^{\circ,p} \in \mathcal{B}^{m,p}(\mathbb{R}^d)$ such that

$$
g_{t,j}^{\circ} = \bar{g}_{t,j}^{\circ,p} |x|
$$

The same result holds if $p = 1$ and $m \geq 2$ is even.

- The case $p = 1$ allows one to apply a result obtained in Chapter 2 and prove the following result.
Proposition 3.5.1

Let Assumption 3.4.1 hold with \( m > d + 1 \) if \( d \) even or \( m > d \) if \( d \) odd. For \( t = 0, \ldots, N - 1 \) and \( j = 1, \ldots, d \) there exists \( C_{t,j} > 0 \) such that for every positive integer \( n_{t,j} \) there exist \( \tau_{t,j,1}, \ldots, \tau_{t,j,n_{t,j}} \in \mathbb{R}^d \), \( b_{t,j,1}, \ldots, b_{t,j,n_{t,j}} > 0 \), \( c_{t,j,1}, \ldots, c_{t,j,n_{t,j}} \in \mathbb{R} \) such that

\[
\sup_{x \in X} \left| g_{t,j}^o(x) - \sum_{i=1}^{n_{t,j}} c_{t,j,i} \exp \left( -\frac{\|x - \tau_{t,j,i}\|^2}{b_{t,j,i}} \right) \right| \leq \frac{C_{t,j}}{\sqrt{n_{t,j}}}.
\]
Error propagation

We combine this result with an estimate of the modulus of continuity of the functional at \((g_0^o, \ldots, g_{N-1}^o)\):

**Theorem 3.5.2**

**Assumptions (p. 44)** ⇒ *There exist suboptimal policies \(\tilde{g}_0, \ldots, \tilde{g}_{N-1}\), with sparse components of the form*

\[
\tilde{g}_{t,j}(x) = \sum_{i=1}^{n_{t,j}} c_{t,j,i} \exp \left( -\frac{\|x - \tau_{t,j,i}\|^2}{b_{t,j,i}} \right),
\]

*such that*

\[
\sup_{x \in X} \left| J^o(x) - \tilde{J}^o(x) \right| \leq \sum_{t=0}^{N-1} \Lambda_t \sqrt{\sum_{j=1}^{d} \frac{C_{t,j}^2}{n_{t,j}}}.
\]

See p. 44 for expressions of the \(\Lambda_t > 0\) and p. 37 for the expression of \(\tilde{J}^o\).
Optimal-policies approximation (V)

- For problems obtained by smooth perturbations of a problem with known optimal policies: similar estimates when only the perturbations of the optimal policies are approximated.
- \( M \): upper bound on the number of computational units, i.e., on \( \sum_{t=0}^{N-1} \sum_{j=1}^{d} n_{t,j} \).
- For smooth perturbations of the linear-quadratic (LQ) regulator problem, we estimate \( \Lambda_t \) and \( C_{t,j} \).
- Optimal allocation of the computational units:

\[
\arg\min_{n_{t,j} \in \mathbb{N}^*} \sum_{t=0}^{N-1} \Lambda_t \sqrt{\sum_{j=1}^{d} \frac{C_{t,j}^2}{n_{t,j}}},
\]

s. t. \( \sum_{t=0}^{N-1} \sum_{j=1}^{d} n_{t,j} \leq M \).
Our estimates of $C_{t,j}$ are defined apart a common factor, which does not influence the optimal allocation of the computational units.

Numerical example: perturbation of a scalar LQ regulator two-stage problem:

$$\arg\min_{n_0, n_1 \in \mathbb{N}^*} \left( \frac{2.36}{\sqrt{n_0}} + \frac{1.28}{\sqrt{n_1}} \right),$$

s.t. $n_0 + n_1 \leq M$.

At optimality one has

$$n_0^o + n_1^o = M.$$ 

Let

$$f(n_0, M) := \left( \frac{2.36}{\sqrt{n_0}} + \frac{1.28}{\sqrt{M - n_0}} \right).$$
A numerical example (II)
Value-functions approximation in Dynamic Programming (I)

Dynamic programming \((T: \text{Bellman’s operator}, \text{p. 35})\)

\[
\begin{align*}
J_N^0(x_N) &= h_N(x_N), \\
J_t^0(x_t) &= (TJ_{t+1}^0)(x_t), \quad t = N - 1, \ldots, 0.
\end{align*}
\]

\(J_t^0: t\)-th value function.

Theorem 4.3.2

Assumptions (p. 65) \(\Rightarrow\) For every \(t = 0, \ldots, N\) and \(1 \leq p \leq \infty\), the following hold:

(i) \(J_t^0 \in \mathcal{W}^m,p(X)\);

(ii) there exists a function \(\tilde{J}_t^0,p \in \mathcal{W}^m,p(\mathbb{R}^d)\) such that \(J_t^0 = \tilde{J}_t^0,p|_X\).

- We investigate the error propagation when \(\hat{J}_t^0(x_t) = (T\tilde{J}_{t+1}^0)(x_t)\) is computed instead of \(J_t^0(x_t)\), and \(\hat{J}_t^0\) is approximated by \(\tilde{J}_t^0\) belonging to a suitable vbs.
Proposition 4.3.3

Assumptions (p. 65) ⇒ For \( t = 0, \ldots, N - 1 \):

(i) If \( \hat{J}^{o}_{t+1} \in C^{m}(X) \) is \( \bar{\alpha}_{t+1} \)-concave for some \( \bar{\alpha}_{t+1} > 0 \), then there exists \( \hat{J}^{o}_{t} \in \mathcal{W}^{m,p}(\mathbb{R}^d) \) such that \( T\hat{J}^{o}_{t+1} = \hat{J}^{o}_{t}|_{X} \).

(ii) Let \( \bar{J}^{o}_{t} \) be defined as in the statement of Theorem 4.3.2 (ii) and for \( j = 1, \ldots, \) let \( \tilde{J}^{o}_{t+1,j} \in C^{m}(X) \) be such that

\[
\lim_{j \to \infty} \max_{0 \leq |r| \leq m} \left\{ \sup_{x \in X} |D^{r}(J^{o}_{t+1}(x) - \tilde{J}^{o}_{t+1,j}(x))| \right\} = 0.
\]

Then, for all sufficiently large \( j \) (i) holds for \( \hat{J}^{o}_{t,j} \) and

\[
\lim_{j \to \infty} \|\bar{J}^{o}_{t} - \hat{J}^{o}_{t,j}\|_{\mathcal{W}^{m,p}(\mathbb{R}^d)} = 0.
\]
Vbs for value-functions approximation

\[ \mathcal{F}(\psi, n) = \left\{ f_n : X \subset \mathbb{R}^d \to \mathbb{R} \mid f_n(x) = \sum_{i=1}^{n} c_i \psi(a_i \cdot x + b_i), a_i \in \mathbb{R}^d, c_i, b_i \in \mathbb{R} \right\}, \]

where \( \psi \) belongs to

\[ S_m := \left\{ \psi : \mathbb{R} \to \mathbb{R} \mid \text{nonzero, compactly supported, with continuous and uniformly bounded partial derivatives up to the order } m, \right\} \]

and \( \exists l \geq m \text{ s.t. } 0 < \int_{\mathbb{R}} |D^l \psi(z)| \, dz < \infty \} . \]

E.g.: splines of suitable order.

\[ S' := \left\{ \psi : \mathbb{R} \to \mathbb{R} \mid \text{nonzero, infinitely many times differentiable on some open interval } (a, b) \subset \mathbb{R}, \text{ and such that } \exists c \in (a, b) : D^k \psi(c) \neq 0 \ \forall k \in \mathbb{N} \right\}. \]

E.g.: Gaussian.
Error propagation (I)

- We consider the approximation capability of the vbs $F(\psi_t, n_t)$.

- Assumptions:
  - $\hat{J}_t^o$ is known exactly (infinite number of samples).
  - $\tilde{J}_t^o \in F(\psi_t, n_t)$ has to be found only on the basis of the available information ($\hat{J}_t^o$, not $J_t^o$, since this is unknown).
Proposition 4.2.2

**Assumptions (p. 61):**

\[
\begin{align*}
\downarrow \\
(i) \text{ If for } t = 0, 1, \ldots, N - 1 \text{ there exists } f_t \in \mathcal{F}(\psi_t, n_t) \text{ such that } \\
\sup_{x \in X} |(T \tilde{J}^0_{t+1}(x) - f_t(x))| \leq \varepsilon_t,
\end{align*}
\]

and one takes \( \tilde{J}^0_t = f_t \), then

\[
\sup_{x \in X} |J^0_0(x) - \tilde{J}^0_0(x)| \leq \sum_{t=0}^{N-1} \beta^t \varepsilon_t.
\]

(ii) If for every \( t = 0, 1, \ldots, N - 1 \) there exists \( f_t \in \mathcal{F}(\psi_t, n_t) \) such that

\[
\sup_{x \in X} |J^0_t(x) - f_t(x)| \leq \bar{\varepsilon}_t,
\]

then one can choose each \( \tilde{J}^0_t \), on the basis of the available information \( \hat{J}^0_t \), such that

\[
\sup_{x \in X} |J^0_0(x) - \tilde{J}^0_0(x)| \leq \sum_{t=0}^{N-1} (2 \beta)^t \bar{\varepsilon}_t.
\]
Error propagation (III)

Under a slight variation of Assumption 3.4.1,

- we derive a-priori upper bounds of the form \( \frac{C_t}{\sqrt{nt}} \)
  
  - on \( \varepsilon_t \), when \( m \) grows linearly in \( d \cdot N \) and \( \psi_t \in S_{2+(2s+1)(t+1)} \), where \( s := \lfloor \frac{d}{2} \rfloor + 1 \) (Proposition 4.4.2 (i));
  
  - on \( \bar{\varepsilon}_t \), when \( m \) (the degree of smoothness in the problem formulation) grows linearly in \( d \) and \( \psi_t \in S' \) (Proposition 4.4.2 (ii));

- we estimate rates of approximate optimization by multistage lookahead variable-basis approximation of the \( t \)-th value functions.
  (Multistage lookahead = iterated application of Bellman’s operator: for a positive integer \( L \), the functions \( \hat{J}_t^o(L) := T(L)\hat{J}_{t+1}^o \) are approximated, instead of the functions \( \hat{J}_t^o \)).
Static stochastic team optimization problems (I)

- **Static team of** $n$ **decision makers (DMs),** $i = 1, \ldots, n$. 
- $x \in X \subseteq \mathbb{R}^d$: random variable, *(state of the world)*, with a probability density $p : X \rightarrow \mathbb{R}$ describing a stochastic environment.
- $y_i = f_i(x) \in Y_i \subseteq \mathbb{R}^{d_i}$: *information* that the DM $i$ has about $x$, which is a given function of the state of the world.
- $s_i : Y_i \rightarrow A_i \subseteq \mathbb{R}$: *strategy* of the $i$-th DM.
- $a_i = s_i(y_i)$: *action* that the DM $i$ chooses on the basis of the information $y_i$.
- $u : X \times \prod_{i=1}^n Y_i \times \prod_{i=1}^n A_i \rightarrow \mathbb{R}$: real-valued *team utility function*, to be jointly maximized by the team.

**Statistical probability structure**: the information that the $n$ DMs have on the state of the world $x$ is modelled by an $n$-tuple of random variables $y_1, \ldots, y_n$, associated (together with $x$) with the joint probability density $q(x, y_1, \ldots, y_n)$. 
Static stochastic team optimization problems (II)

Problem STO (Static team optimization with statistical information).

Given the random variable $x$ with a probability density $p : X \to \mathbb{R}$, the statistical information structure $q(x, y_1, \ldots, y_n)$ having $p(x)$ as its marginal density on $X$, the team utility function

$$ u(x, y_1, \ldots, y_n, a_1, \ldots, a_n), $$

$$ u : \mathbb{R}^N \to \mathbb{R}, $$

where $N = d + \sum_{i=1}^n d_i + n$, find

$$ \sup_{s_1, \ldots, s_n} v(s_1, \ldots, s_n), $$

where

$$ v(s_1, \ldots, s_n) = \mathbb{E}_{x, y_1, \ldots, y_n} \left\{ u(x, \{y_i\}_{i=1}^n, \{s_i(y_i)\}_{i=1}^n) \right\}. $$

$$ \sup_{s_1, \ldots, s_n} v(s_1, \ldots, s_n): \text{value of the team}. $$
Concavity at least $\tau$

- A concave function $f$ defined on a convex set $X$ has concavity (at least) $\tau > 0$ if for all $u, v \in X$ and every supergradient $a_u$ of $f$ at $u$ one has $f(v) - f(u) \leq a_u \cdot (v - u) - \tau\|v - u\|^2$.

- If a function has concavity $\tau > 0$, then it is $\tau$-concave.
Assumptions

Assumption 5.2.1

The set $X$ of the states of the world is compact, $Y_1, \ldots, Y_n$ are compact and convex, and $A_1, \ldots, A_n$ are bounded closed intervals. For an integer $m \geq 2$, the utility $u$ is of class $C^m$ on an open set containing $X \times \prod_{i=1}^{n} Y_i \times \prod_{i=1}^{n} A_i$ and $q$ is a (strictly) positive probability density on $X \times \prod_{i=1}^{n} Y_i$, which can be extended to a function of class $C^m$ on an open set containing $X \times \prod_{i=1}^{n} Y_i$.

Assumption 5.2.2

There exists $\tau > 0$ such that the team utility function $u : \mathbb{R}^N \to \mathbb{R}$ is separately concave in each of the decision variables $s_1, \ldots, s_n$, with concavity $\tau$ in each of them.
Smoothness of the optimal team strategies

Theorem 5.3.5

Let Assumptions 5.2.1 and 5.2.2 hold. If for every n-tuple \( \{s_1, \ldots, s_n\} \) of strategies, the strategies defined as

\[
\hat{s}_1(y_1) := \arg\max_{a_1 \in A_1} E_{x_1, y_2, \ldots, y_n} \{ u(x, \{y_i\}_{i=1}^n, a_1, \{s_i(y_i)\}_{i=2}^n) \},
\]

\[
\vdots
\]

\[
\hat{s}_n(y_n) := \arg\max_{a_n \in A_n} E_{x_n, y_1, \ldots, y_{n-1}} \{ u(x, \{y_i\}_{i=1}^n, \{s_i(y_i)\}_{i=1}^{n-1}, a_n) \}
\]

do not lie on the boundaries of \( A_1, \ldots, A_n \), respectively, then

Problem STO admits \( C^{m-2} \) optimal strategies \( (s_1^o, \ldots, s_n^o) \) with partial derivatives that are Lipschitz up to the order \( m - 2 \).
Proposition 5.4.3

Let the assumptions of Theorem 5.3.5 hold with an odd integer $m > \max_i \{d_i\} + 1$. Then there exist $K(m, d_i) > 0$, $i = 1, \ldots, n$, and a positive constant $C$, which depends on $(s^o_1, \ldots, s^o_n)$, such that for every positive integer $k$ there exists an $n$-tuple of strategies $(\tilde{s}^k_1, \ldots, \tilde{s}^k_n)$ such that

$$v(s^o_1, \ldots, s^o_n) - v(\tilde{s}^k_1, \ldots, \tilde{s}^k_n) \leq \frac{C}{\sqrt{k}},$$

$$\tilde{s}^k_i(y_i) = \sum_{j=1}^{k} c_{ij} \text{Prj}_{A_i}(g_{ij}(y_i)), \ g_{ij} \in G_i,$$

$$G_i = \left\{ g_i : Y_i \to \mathbb{R} \mid g_i(y_i) = e^{-\frac{\|y_i - t_i\|^2}{\delta_i}}, t_i \in \mathbb{R}^{d_i}, \delta_i > 0 \right\},$$

$$\sum_{j=1}^{k} |c_{ij}| \leq K(m, d_i)\|\lambda_i\|_{L^1(\mathbb{R}^{d_i})},$$

and $\lambda_i \in L^1(\mathbb{R}^{d_i})$ is such that $s^o_i,\text{ext},^1 = G_{m-1} * \lambda_i$ for a suitable extension $s^o_i,\text{ext},^1$ of $s^o_i$ on $\mathbb{R}^{d_i}$.
For some vbs with \( k \) computational units for each DM: upper bounds of approximate optimization of order \( \frac{1}{\sqrt{k}} \), for a degree of smoothness \( m \) in the problem formulation that grows linearly in \( \max_{i=1}^{n} d_i \).

More decentralization \( \rightarrow \) Lower degree of smoothness required.

The estimates may be extended to dynamic team optimization problems, as several dynamic team optimization problems are equivalent to static ones (Witsenhausen, 1988).
Sparse suboptimal solutions in learning from data

For weight decay and Tikhonov, Ivanov, Miller, and Phillips regularizations:

- **Representer Theorem** in Reproducing Kernel Hilbert Spaces.
- Estimates of the *moduli of continuity* of the objective functionals at the optimal solutions.
- **Rates of approximate optimization** by sparse suboptimal solutions with kernel computational units.

- Estimates improved in Gnecco and Sanguineti, “Regularization techniques and suboptimal solutions to optimization problems in learning from data”, submitted.

- Improvements of Theorem 6.5.2, 6.6.4, and 6.6.5.

- **Bounds from Statistical Learning Theory** on the difference between the expected error and the empirical error for sparse suboptimal solutions.
Reproducing Kernel Hilbert Spaces (RKHSs)

- **Real RKHS** ($\mathcal{H}_K(X)$, $\| \cdot \|_{\mathcal{H}_K(X)}$): real Hilbert space formed by functions defined on a non-empty set $X$ such that for every $u \in X$, the evaluation functional $\mathcal{F}_u$

  \[(\mathcal{F}_u(f) := f(u))\]

  is bounded.

- RKHSs can be also characterized in terms of **symmetric positive-semidefinite kernels**, which are symmetric functions $K : X \times X \rightarrow \mathbb{R}$ such that, for every $m \in \mathbb{N}^*$ and every $x = (x_1, \ldots, x_m) \in X^m$, the $m \times m$ **Gram matrix** $\mathcal{K}[x]$, with entries

  \[\mathcal{K}[x]_{i,j} := K(x_i, x_j),\]

  is positive-semidefinite.
Vbs for Ivanov regularization in learning from data

- For \( i = 1, \ldots, m \): \((x_i, y_i) \in X \times Y\), where \( X \subseteq \mathbb{R}^d \) and \( Y \subseteq \mathbb{R} \) (training set).
- \((\mathcal{H}_K(X), \| \cdot \|_{\mathcal{H}_K(X)})\): (real) Reproducing Kernel Hilbert Space on \( X \), associated with the symmetric positive semi-definite kernel \( K : X \times X \to \mathbb{R} \).
- \( r > 0 \): regularization parameter.
- \( M_I(r) := \{ f \in \mathcal{H}_K(X) : \| f \|_{\mathcal{H}_K(X)} \leq r \} \): hypothesis set.

Vbs for Ivanov regularization \((n < m)\):

**Theorem 6.5.4 Assumptions (p. 113)**

\[ \downarrow \]

\((i)\) \( \inf_{f \in M_I(r) \cap \text{span}_n G_{K_X}} \Phi_{I,r}(f) - \Phi_{I,r}(f_{I,r}^o) \leq a_1 \sqrt{\frac{\Delta_{I,r}}{n}} + a_2 \frac{\Delta_{I,r}}{n} \).

\((ii)\) \( \varepsilon_n \geq 0 \), \( f_n \) an \( \varepsilon_n \)-near minimum point of \((M_I(r) \cap \text{span}_n G_{K_X}, \Phi_{I,r})\):

\[ \| f_n - f_{I,r}^o \|_{\mathcal{H}_K(X)}^2 \leq \frac{1}{\gamma(r)} \left( a_1 \sqrt{\frac{\Delta_{I,r}}{n}} + a_2 \frac{\Delta_{I,r}}{n} + \varepsilon_n \right) + \left( \| f_n \|_{\mathcal{H}_K(X)}^2 - r^2 \right). \]
## Regularization techniques for learning from data in RKHSs

<table>
<thead>
<tr>
<th>Regularization technique</th>
<th>Regularization parameter</th>
<th>Functional</th>
<th>Hypothesis set</th>
<th>Minimization problem</th>
<th>Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tikhonov</td>
<td>$\gamma &gt; 0$</td>
<td>$\Phi_{T,\gamma} = \mathcal{E}<em>z + \gamma | \cdot |^2</em>{\mathcal{H}_K(X)}$</td>
<td>$\mathcal{H}_K(X)$</td>
<td>Problem $T_{\gamma}$: $(\mathcal{H}_K(X), \mathcal{E}<em>z + \gamma | \cdot |^2</em>{\mathcal{H}_K(X)})$</td>
<td>$f^\circ_{T,\gamma}$</td>
</tr>
<tr>
<td>Ivanov</td>
<td>$r &gt; 0$</td>
<td>$\Phi_{I,r} = \mathcal{E}_z$</td>
<td>$B_r(| \cdot |_{\mathcal{H}_K(X)})$</td>
<td>Problem $I_r$: $(B_r(| \cdot |_{\mathcal{H}_K(X)}), \mathcal{E}_z)$</td>
<td>$f^\circ_{I,r}$</td>
</tr>
<tr>
<td>Phillips</td>
<td>$\eta &gt; 0$</td>
<td>$\Phi_{P,\eta} = | \cdot |^2_{\mathcal{H}_K(X)}$</td>
<td>$F_{\eta}$</td>
<td>Problem $P_{\eta}$: $(F_{\eta}, | \cdot |^2_{\mathcal{H}_K(X)})$</td>
<td>$f^\circ_{P,\eta}$</td>
</tr>
<tr>
<td>Miller</td>
<td>$r, \eta &gt; 0$</td>
<td>$\Phi_{M,r,\eta} = \mathcal{E}<em>z + (\frac{\eta}{r})^2 | \cdot |^2</em>{\mathcal{H}_K(X)}$</td>
<td>$G_{r,\eta}$</td>
<td>Problem $M_{r,\eta}$: $(G_{r,\eta}, \mathcal{E}<em>z + (\frac{\eta}{r})^2 | \cdot |^2</em>{\mathcal{H}_K(X)})$</td>
<td>$f^\circ_{M,r,\eta}$</td>
</tr>
</tbody>
</table>
**Weight-decay, Tikhonov regularization, and mixed weight-decay/Tikhonov learning**

<table>
<thead>
<tr>
<th>Learning technique</th>
<th>Functional</th>
<th>Minimization problem</th>
<th>Linear system of equations</th>
<th>Spectral window</th>
</tr>
</thead>
<tbody>
<tr>
<td>WD</td>
<td>$\Phi_{\text{WD}}, \gamma(f) = \varepsilon(f) + \gamma | c_f |_2^2$</td>
<td>$(\text{span}<em>m G_K, \Phi</em>{\text{WD}}, \gamma)$</td>
<td>$(\mathcal{K}[x] + \gamma , m \mathcal{K}^{-1}[x]) , c = y$</td>
<td>$\frac{\lambda}{\lambda + \gamma , m \lambda - I}$</td>
</tr>
<tr>
<td>Tikhonov</td>
<td>$\Phi_{\text{T}}, \gamma(f) = \varepsilon(f) + \gamma | f |_2^2 \mathcal{H}_K(X)$</td>
<td>$(\mathcal{H}<em>K(X), \Phi</em>{\text{T}}, \gamma)$</td>
<td>$(\gamma , m , I + \mathcal{K}[x]) , c = y$</td>
<td>$\frac{\lambda}{\lambda + \gamma , m}$</td>
</tr>
<tr>
<td>WD/Tikhonov</td>
<td>$\Phi_{\text{WDT}}, \frac{\gamma}{2}(f) = \varepsilon(f) + \frac{\gamma}{2} \left( | c_f |_2^2 + | f |_2^2 \mathcal{H}_K(X) \right)$</td>
<td>$(\text{span}<em>m G_K, \Phi</em>{\text{WDT}}, \frac{\gamma}{2})$</td>
<td>$(\mathcal{K}[x] + \frac{\gamma}{2} , m ,(I + \mathcal{K}^{-1}[x])) , c = y$</td>
<td>$\frac{\lambda}{\lambda + \frac{\gamma}{2} , m(1+\lambda - I)}$</td>
</tr>
</tbody>
</table>
Estimates of $G$-variation (I)

Some upper bounds in Chapters 5 and 6 are based on MJB’s lemma, restated here in terms of $G$-variation:

**Lemma**

For a bounded subset $G$ of a Hilbert space $(\mathcal{H}, \| \cdot \|_\mathcal{H})$ with $s_G = \sup_{g \in G} \| g \|_\mathcal{H}$ and every $f \in \mathcal{H}$:

$$\| f - \text{span}_n G \|_\mathcal{H} \leq \sqrt{\frac{(s_G \| f \|_G)^2 - \| f \|_\mathcal{H}^2}{n}}.$$ 

where

$$\| f \|_G := \inf\{ c > 0 : f/c \in \text{cl conv}(G \cup -G) \}.$$

In Chapter 7 we derive upper and lower bounds on $G$-variation in terms of “classical” norms, for functions $f$ with the integral representation

$$f(x) = \int_X w(y)K(x, y)dy$$

$(X \subset \mathbb{R}^d, K : X \times X \to \mathbb{R}$ continuous, and $w \in L^2$ with compact support $A)$.
Estimates of $G$-variation (II)

Theorem 7.2.3

Let $k, w : \mathbb{R}^d \to \mathbb{R}$ be continuous, $w \in \mathcal{L}^2$ have compact support $A \subset \mathbb{R}^d$, $k \in \mathcal{L}^2(\mathbb{R}^d)$ be an even function such that for all $s \in \mathbb{R}^d$ its Fourier transform $\hat{k}(s) \geq 0$ and is continuous, $f = w \ast k$, and $h = k \ast k$. Then

$$\|f\|_{G_K} \leq \frac{\sqrt{\lambda(A)}}{(2\pi)^{d/4}} \|f\|_{\mathcal{H}_H},$$

where the closure in the definition of $G_K$-variation is with respect to the $\mathcal{L}^2(\mathbb{R}^d)$-norm.

$\mathcal{H}_H$: (real) Reproducing Kernel Hilbert Space on $\mathbb{R}^d$, with translation-invariant kernel $H(x, y) = h(x - y)$.

For several kernels $K$, we exploit this result to derive upper bounds on $G_K$-variation (associated to the set $G_K := \{K(x, \cdot)\}$).
## Estimates of $G$-variation (III)

<table>
<thead>
<tr>
<th>Kernel</th>
<th>Set of functions</th>
<th>Upper bound</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gaussian:</strong> ( \Phi_b(x, y) = e^{-b|x-y|^2} )</td>
<td>( \Theta_{\varphi_b,A} = { f = w * \varphi_b } )</td>
<td>( | f |<em>{G</em>{\Phi_b}} \leq \frac{2d^2b}{d(2\pi)^d} \sqrt{\lambda(A)} | f |<em>{H</em>{\varphi_b/2}} )</td>
</tr>
<tr>
<td><strong>Laplace:</strong> ( \Psi_a(x, y) = e^{-a|x-y|} )</td>
<td>( \Theta_{\psi_a,A} = { f = w * \psi_a } )</td>
<td>( | f |<em>{G</em>{\Psi_a}} \leq \sqrt{2} \sqrt{\lambda(A)} \times \sqrt{| f |^2_{L^2} + \frac{2}{a} | f' |^2_{L^2} + \frac{1}{a^2} | f'' |^2_{L^2}} )</td>
</tr>
<tr>
<td><strong>Bessel:</strong> ( G_r(x - y), \text{s.t.} ) ( \hat{G}_r(s) = \frac{1}{(1+|s|^2)^{r/2}} )</td>
<td>( \Theta_{G_r,A} = { f = w * G_r } )</td>
<td>( | f |<em>{G</em>{G_r}} \leq \sqrt{\lambda(A)} | f |_{B_r,2} )</td>
</tr>
<tr>
<td><strong>Sinc:</strong> ( \eta_b(x - y), \text{s.t.} ) ( \hat{\eta}_b(s) = \frac{1}{\sqrt{2\pi b^2}} \text{rect} \left( \frac{s}{2\pi b} \right) )</td>
<td>( \Theta_{\eta_b,A} = { f = w * \eta_b } )</td>
<td>( | f |<em>{G</em>{H_b}} \leq \sqrt{\lambda(A)} b \sqrt{2\pi} | f |_{L^2} )</td>
</tr>
</tbody>
</table>
Estimates of $G$-variation (IV)

For some kernels, we also obtain a lower bound on $G_K$-variation. This, combined with a previously available upper bound, gives the following

**Proposition 7.4.1**

Let $X \subset \mathbb{R}^d$ be compact, $f : X \to \mathbb{R}$ a continuous function such that $f(x) = \int_X w(y)K(x, y) \, dy$, $w : X \to \mathbb{R}$ continuous and non-negative, $K : X \times X \to \mathbb{R}$ symmetric and continuous, and assume that constant functions belong to the range of the integral operator $T : \mathcal{L}^2(X) \to \mathcal{L}^2(X)$ defined as $(T g)(x) = \int_X g(y)K(x, y) \, dy$. Then

$$\|f\|_{G_K} = \|w\|_{\mathcal{L}^1(X)},$$

where the closure in the definition of $G_K$-variation is with respect to the $\mathcal{L}^2(X)$-norm.

For $d = 1$, an extension to weight functions $w$ without the non-negativity constraint is given in Proposition 7.4.2.
Current developments

- **Dynamic optimization problems:**
  - Extensions of the results to the *stochastic case* (optimal consumption under uncertainty).
  - Extensions of the results to *other models* (state equations and control constraints instead of correspondences):
    - control-affine dynamic systems.
    - dynamic problems obtained as perturbations of static ones.

- **Learning from data:**
  - Improved estimates.
  - **Statistical Learning Theory bounds** for sparse suboptimal solutions.

- **Algorithmic issues** (Information-based complexity results for the ERIM).
Estimates of the absolute constants in the error bounds.

**Dynamic optimization** problems:
- Extensions to infinite-horizon problems (beside local results).
- Extensions to nonconcave problems.
- Extensions to continuous-time problems.

**Team optimization** problems:
- Extensions of the results to dynamic problems.
- Extensions of the results to game theory problems.

**Vbs** for integral equations.

**Vbs** for learning and optimization problems on graphs.

Greedy algorithms.

Error bounds for vbs approximation of families of nonsmooth functions.
Thanks for the attention!
Difficulties in finding optimal solutions: an example (I)

Tikhonov regularization in supervised learning from data:

- For \( i = 1, \ldots, m \): \((x_i, y_i) \in X \times Y\), where \( X \subseteq \mathbb{R}^d \) and \( Y \subseteq \mathbb{R} \) (training set).
- \((\mathcal{H}_K(X), \| \cdot \|_{\mathcal{H}_K(X)})\): (real) Reproducing Kernel Hilbert Space on \( X \), associated with the symmetric positive semi-definite kernel \( K : X \times X \to \mathbb{R} \).
- \( \gamma > 0 \): regularization parameter.

Problem \((\mathcal{H}_K(X), \Phi_T, \gamma)\):

\[
\inf_{f \in \mathcal{H}_K(X)} \left( \frac{1}{m} \sum_{i=1}^{m} (y_i - f(x_i))^2 + \gamma \| f \|_{\mathcal{H}_K(X)}^2 \right).
\]
Difficulties in finding optimal solutions: an example (II)

- $\mathcal{K}[\mathbf{x}]$: $(m \times m)$ Gram matrix of the data, $\mathcal{K}[\mathbf{x}]_{i,j} = K(x_i, x_j)$.
- $\mathbf{y} = (y_1, \ldots, y_m)^T$.

**Representer Theorem for Tikhonov-regularized learning**

(e.g., Cucker and Smale, 2002) The optimal solution to Problem $(\mathcal{H}_\mathcal{K}(\mathcal{X}), \Phi_T, \gamma)$ is

$$f^o(\cdot) = \sum_{i=1}^{m} c_i^o K(x_i, \cdot),$$

where $\mathbf{c} = (c_1^o, \ldots, c_m^o)^T$ is the unique solution to the well-posed linear system

$$(\gamma m \mathcal{I} + \mathcal{K}[\mathbf{x}])\mathbf{c}^o = \mathbf{y}.$$
Difficulties in finding optimal solutions: an example (III)

Drawbacks of the Representer Theorem:

- Usually, the matrix $\mathcal{K}[x]$ is not sparse. Finding $f^o$ requires one to invert a nonsparse $m \times m$ matrix: this is a difficult problem for large $m$.

- For large sample size $m$: usually, many nonzero coefficients $c_i^o$ in the expression of $f^o$.

...Interest in deriving sparse suboptimal solutions (with a much smaller number of nonzero coefficients).
Why looking for sparse suboptimal solutions?

- **Storing** the coefficients of a sparse suboptimal solution $f^s$ requires less memory than for $f^o$.
- The **time** required to find $f^s$ may be much smaller than for $f^o$. E.g.: greedy algorithms (Zhang, 2002).
- Once $f^s$ has been found, computing $f^s(x)$ for $x$ not belonging to the input-data set requires less time than computing $f^o(x)$.
- A sparse model has a better interpretability than a non-sparse one (it is a simpler model, which depends on less parameters). E.g.: gene selection problem in microarray data analysis (Zou and Hastie, 2005).
The curse of dimensionality in the number of basis functions

For certain function approximation problems in real Hilbert spaces \((\mathcal{H}, \| \cdot \|_\mathcal{H})\), for which

- \(S = \mathcal{H}\),
- \(\bar{f} : X \subseteq \mathbb{R}^d\) is a function belonging to a suitable subset of \(\mathcal{H}\),
- \(\Phi(f) = \| f - \bar{f} \|_\mathcal{H}\),
- \(\inf_{f \in S} \Phi(f) = 0\),

(Barron, 1993) showed that the best fixed-basis approximation schemes suffer from the curse of dimensionality in \(n\), while for certain vbs

\[
\inf_{f \in \text{span}_n G} \| f - \bar{f} \|_\mathcal{H} \leq C(d) \sqrt{\frac{1}{n}},
\]

where (at least) the second factor does not depend on \(d\).

Classes of problems for which \(C(d)\) grows “slowly” with \(d\) (Kainen, Kůrková, & Sanguineti, 2008).
(Zoppoli, Sanguineti, & Parisini, 2002) suggested to use VBS with sigmoidal or radial computational units to find sparse suboptimal solutions to Problem \((S, \Phi)\): **Extended Ritz Method (ERIM)**.

- Suboptimal solutions of this form can be found by solving a finite-dimensional nonlinear programming problem (e.g., training a neural network).
- **Numerical results** (e.g., Zoppoli, Sanguineti, and Parisini, 2002).
- Only a few **theoretical results** (Kůrková and Sanguineti, 2004) on

\[
\inf_{f \in S \cap \text{span}_n G} \Phi(f) - \inf_{f \in S} \Phi(f),
\]

for \(\Phi\) slightly more general than in (Barron, 1993).