GRAN SASSO Science Institute
Intensive Trimester "Particles, Fluids and Patterns: Analytical and Computational Challenges"
3-4 June 2022

High-dimensional approximation and sampling

Part 4: Approximation from limited information

Anthony Nouy

Centrale Nantes, Nantes Université, Laboratoire de Mathématiques Jean Leray

Approximation

We consider the approximation of a function f of a normed space V by an element of a set V_m described by m parameters.

An approximation tool $(V_m)_{m\geq 1}$ is selected from some prior knowledge on the function class K to approximate, for obtaining a fast (hopefully optimal) convergence of the best approximation error

$$\inf_{g\in V_m} \|f-g\|_V$$

- Analytic smoothness: polynomials
- Sobolev or Besov smoothness: splines, wavelets
- For a larger class of functions: tensor networks, neural networks
- Low-dimensional space or manifold $V_m = \{D(a) : a \in \mathbb{R}^m\}$ which approximates K, obtained by manifold approximation methods.

Approximation from limited information

In practice, an approximation in V_m is produced by an algorithm A_n using only a limited number of information $L_1(f), \ldots, L_n(f)$ and returning

$$A_n(f) = R(L_1(f), \ldots, L_n(f))$$

where R is a reconstruction map with values in V_m.

An algorithm is quasi-optimal for a function class if for any function from this class,

$$\|f - A_n(f)\|_V \leq C \inf_{g \in V_m} \|f - g\|_V$$

A random algorithm is quasi-optimal in average (of order p) if

$$\mathbb{E}(\|f - A_n(f)\|_V^p)^{1/p} \leq C \inf_{g \in V_m} \|f - g\|_V$$

Type of information

Different types of information (context dependent)

- pointwise evaluations of the function (aka standard information)

$$L_i(f) = f(x_i)$$

- pointwise evaluations of the function and its derivatives

$$L_i(f) = (D^\alpha f(x_i))_{|\alpha| \leq s}$$

- linear forms

$$L_i(f) = \langle \varphi_i, f \rangle$$

- linear (or nonlinear) maps

$$L_i(f) = \langle \varphi_i, Bf \rangle \quad \text{or} \quad (Bf)(x_i)$$

with B some linear (or nonlinear) operator, e.g. for solving $Bf = g$.

This is the framework of Galerkin or variational methods for PDEs, Physics-informed machine learning (Deep-Galerkin, Deep-Ritz, PINN, ...).

Type of information

We distinguish two different settings:

- Information is given (passive learning). The complexity of the model class V_m is limited. Adaptive strategies play with a collection of model classes $(V_m)_{m\geq 1}$ and require model selection techniques to take the best from the available information.
- Information can be freely generated (active learning). A typical setting in computer/physical experiments, numerical analysis of PDEs, or scientific machine learning.

Optimal information in active learning

A fundamental question is how to generate a good (or optimal) information for a given model class V_m.

When getting information is costly, a challenge is to provide quasi-optimal algorithms using a number of information n close to the number of parameters m.

Adaptive strategies play with a collection of model classes $(V_m)_{m\geq 1}$ and generate information adaptively. A question is then to recycle information in order to obtain a near-optimal performance in terms of information.

Outline

Linear approximation from point evaluations

We consider the approximation of functions from a normed space V of functions defined on a set $\mathcal{X}$, using point evaluations (standard information) and linear algorithms (linear approximation).

We assume that we are given a m-dimensional linear space V_m.

The question is how to generate good points in $\mathcal{X}$ and a linear algorithm that allow to obtain an approximation in V_m with an error close to the best approximation error.

Interpolation

For a set of points $\boldsymbol{x} = (x_1, \ldots, x_m)$ unisolvent for V_m, we let $\mathcal{I}_{V_m} : V \to V_m$ be the corresponding interpolation (linear) operator.

We have

$$\|f - \mathcal{I}_{V_m} f\|_V \leq (1 + \|\mathcal{I}_{V_m}\|) \inf_{v \in V_m} \|f - v\|_V$$

For $(V, \|\cdot\|_\infty)$ the set of functions with bounded norm $\|f\|_\infty := \sup_{x \in \mathcal{X}} |f(x)|$, $\|\mathcal{I}_{V_m}\|$ is the Lebesgue constant, with

$$\|\mathcal{I}_{V_m}\| = \sup_{x \in \mathcal{X}} \sum_{i=1}^{m} |h_i(x)|$$

where $h_1, \ldots, h_m$ is the basis of V_m satisfying the interpolation property ($h_i(x_j) = \delta_{ij}$ for all i, j).

For univariate functions and classical spaces V_m (polynomials, splines), the theory is well established and suitable choices of points are available.

Except in very specific cases (e.g. piecewise constant or linear approximation), $\|\mathcal{I}_{V_m}\|$ grows with m. The question is to find good points such that $\|\mathcal{I}_{V_m}\|$ grows not too fast with m.

Empirical interpolation based on feature map

Let $\boldsymbol{\varphi}(x) = (\varphi_1(x), \ldots, \varphi_m(x)) \in \mathbb{R}^m$, where $\boldsymbol{\varphi} : \mathcal{X} \to \mathbb{R}^m$ is the feature map associated with V_m. The feature space $\mathbb{R}^m$ is equipped with the Euclidian norm $\|\cdot\|$.

The idea is to construct an increasing sequence of spaces

$$U_k = span\{\boldsymbol{\varphi}(x_1), \ldots, \boldsymbol{\varphi}(x_k)\} \subset \mathbb{R}^m$$

for the approximation of the manifold $\{\boldsymbol{\varphi}(x) : x \in \mathcal{X}\}$.

Starting from $U_0 = \{0\}$, we define

$$x_k \in \arg\max_{x \in \mathcal{X}} \Lambda_k(x), \quad \Lambda_k(x) = \|\boldsymbol{\varphi}(x) - P_{U_{k-1}}\boldsymbol{\varphi}(x)\|_2^2$$

where $P_{U_{k-1}}$ is the orthogonal projection from $\mathbb{R}^m$ to U_{k-1}.

Empirical interpolation based on feature map

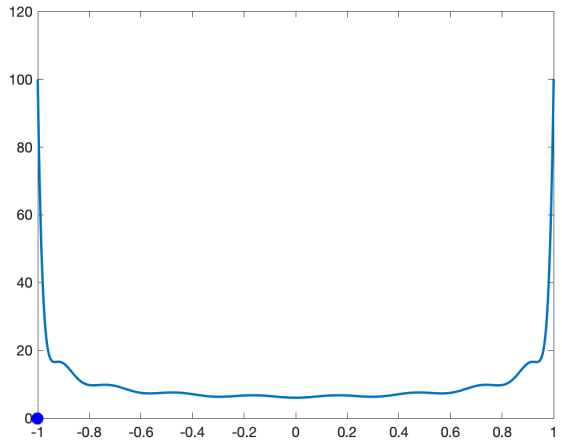
Let $(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_m)$ be the orthonormal basis of $\mathbb{R}^m$ defined by

$$\boldsymbol{e}_k \propto \boldsymbol{\varphi}(x_k) - P_{U_{k-1}} \boldsymbol{\varphi}(x_k), \quad \|\boldsymbol{e}_k\|_2 = 1.$$

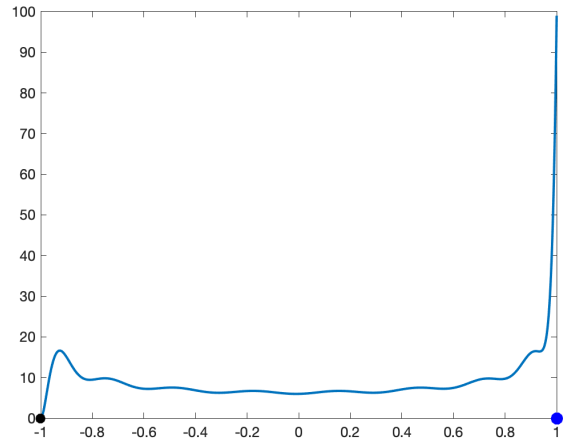
If V_m is a Hilbert space and the functions φ_i form an orthonormal basis of V_m, then the functions $\psi_i(x) = \boldsymbol{\varphi}(x)^T \boldsymbol{e}_i$ also form an orthonormal basis of V_m and

$$\Lambda_k(x) = \sum_{i=k}^{m} \psi_i(x)^2 = \|\boldsymbol{\varphi}(x)\|_2^2 - \sum_{i=1}^{k-1} \psi_i(x)^2$$

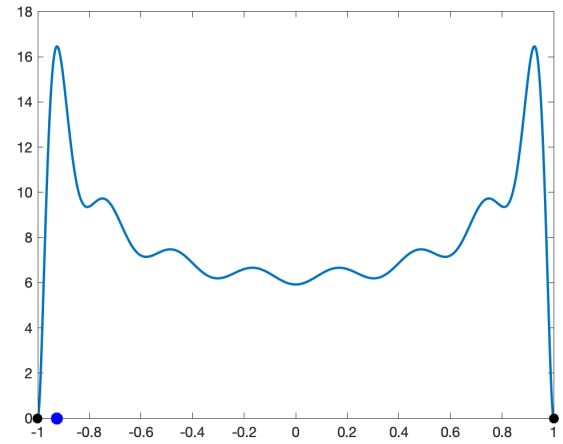
Empirical interpolation based on feature map



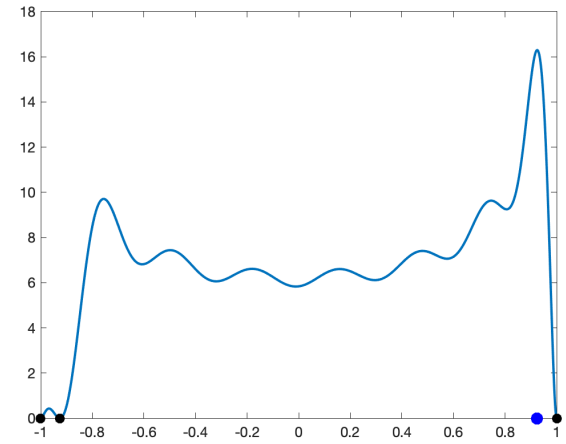
(a) $k = 1$



(b) $k = 2$



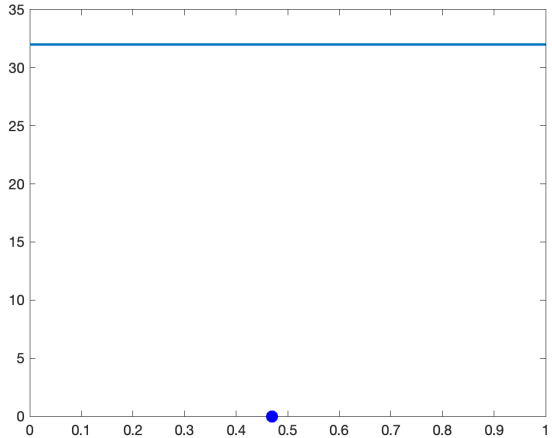
(c) $k = 3$



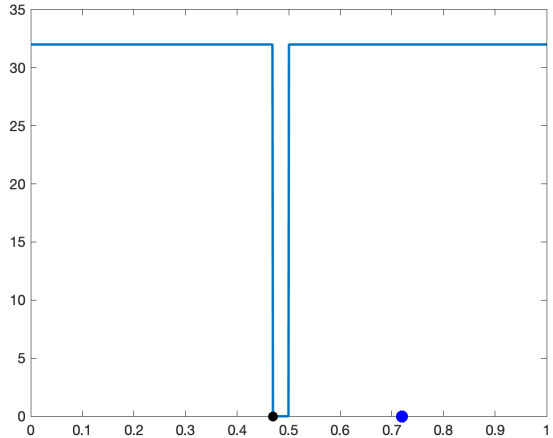
(d) $k = 4$

Figure: Polynomial space $V_m = \mathbb{P}_9$ on $[-1, 1]$. Function $\Lambda_k(x)$ and corresponding interpolation point $x_k = \arg\max_x \Lambda_k(x)$

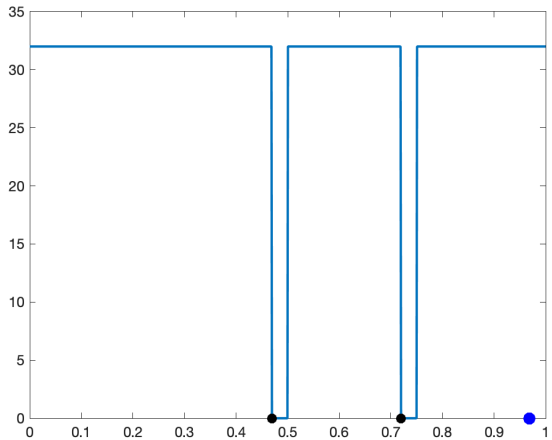
Empirical interpolation based on feature map



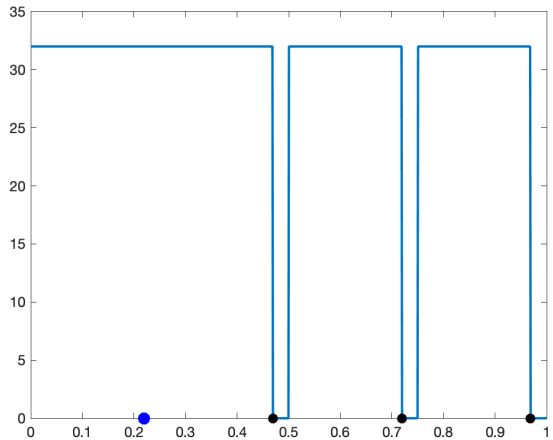
(a) $k = 1$



(b) $k = 2$



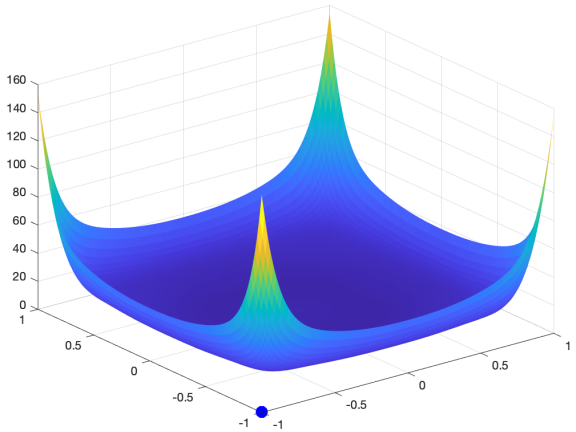
(c) $k = 3$



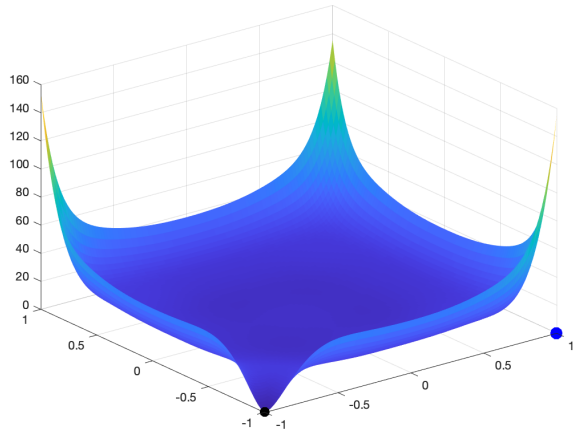
(d) $k = 4$

Figure: Haar wavelets space V_m on $[0, 1]$, with resolution 5. Function $\Lambda_k(x)$ and corresponding interpolation point $x_k = \arg\max_x \Lambda_k(x)$.

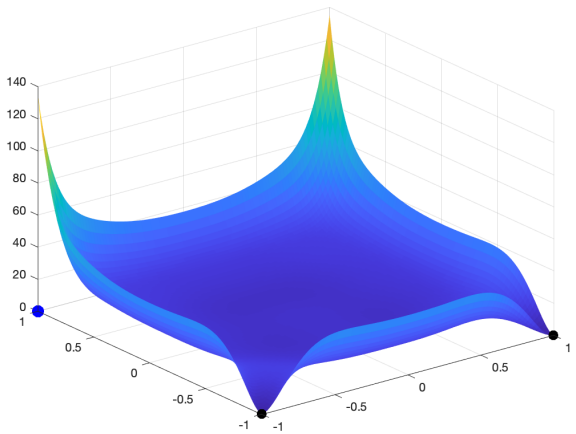
Empirical interpolation based on feature map



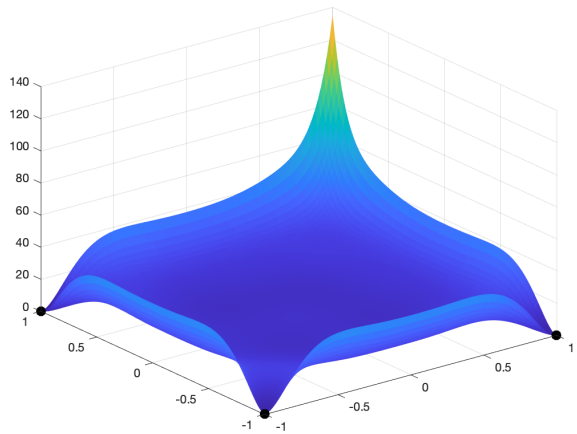
(a) $k = 1$



(b) $k = 2$



(c) $k = 3$



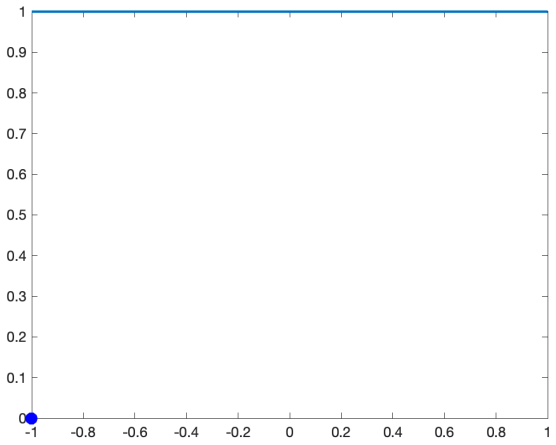
(d) $k = 4$

Figure: Bivariate polynomial space $V_m = \mathbb{P}_4$ on $[-1, 1]^2$. Function $\Lambda_k(x)$ and corresponding interpolation point $x_k = \arg\max_x \Lambda_k(x)$.

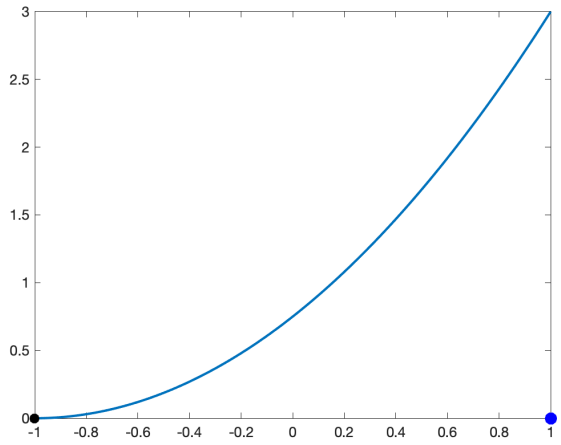
Empirical interpolation based on feature map — adaptive setting

In the context of adaptive approximation in a sequence of spaces $V_1 \subset \ldots \subset V_m \subset \ldots$, and in order to recycle interpolation points, we modify the algorithm by considering at step k the feature map φ associated with the basis of V_k.

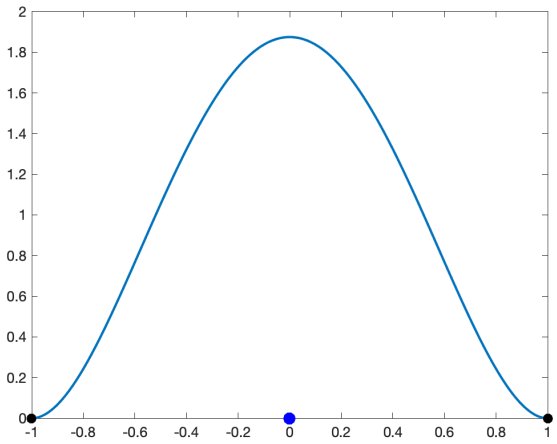
Empirical interpolation based on feature map — adaptive setting



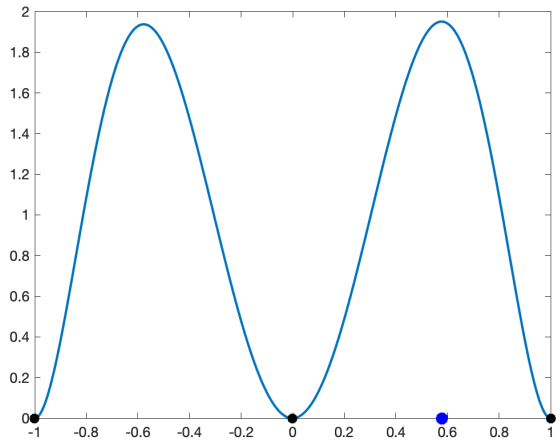
(a) $k = 1$



(b) $k = 2$



(c) $k = 3$



(d) $k = 4$

Figure: Polynomial space $V_m = \mathbb{P}_9$ on [-1,1]. Function $\Lambda_k(x)$ and corresponding interpolation point $x_k = \arg\max_x \Lambda_k(x)$.

Interpolation in RKHS

A reproducing kernel Hilbert space (RKHS) H is a Hilbert space of functions defined on $\mathcal{X}$ such that the point evaluation $\delta_x : f \mapsto f(x)$ is a continuous linear map. There is a so called reproducing kernel k such that $k(x, \cdot)$ is the Riesz representer of δ_x, that is

$$f(x) = (f, k(x, \cdot))_H,$$

where $(\cdot, \cdot)_H$ is the inner product on H.

For given points $\boldsymbol{x} = (x_1, \ldots, x_k)$, the interpolation operator $\mathcal{I}_{W_k}$ onto the space $W_k = span\{k(\cdot, x_1), \ldots, k(\cdot, x_k)\}$ is defined by

$$\mathcal{I}_{W_k} f(x) = k(x, \boldsymbol{x}) k(\boldsymbol{x}, \boldsymbol{x})^{-1} f(\boldsymbol{x})$$

where $k(\boldsymbol{x}, \boldsymbol{y}) = (k(x_i, y_j))_{i,j}$ and $f(\boldsymbol{x}) = (f(x_j))_j$. The operator $\mathcal{I}_{W_k}$ is the

H-orthogonal projection onto W_k, which provides the element of best approximation of a function in W_k. Indeed, for $f \in H$, the interpolation conditions

$$\mathcal{I}_{W_k} f(x_i) = f(x_i), \quad 1 \leq i \leq k,$$

are equivalent to

$$(k(\cdot, x_i), \mathcal{I}_{W_k} f - f)_H = 0, \quad 1 \leq i \leq k,$$

that is $\mathcal{I}_{W_k} f - f$ is orthogonal to W_k.

Interpolation in RKHS

The error of interpolation at point $x \in \mathcal{X}$ is such that

$$\begin{aligned} |f(x) - \mathcal{I}_{W_k} f(x)| &= |\left(k(x,\cdot), \mathcal{I}_{W_k} f - f\right)_H| \\ &= |\left(k(x,\cdot) - \mathcal{I}_{W_k} k(x,\cdot), \mathcal{I}_{W_k} f - f\right)_H| \\ &\leq \|k(x,\cdot) - \mathcal{I}_{W_k} k(x,\cdot)\|_H \|f\|_H \end{aligned}$$

A natural definition of a new basis function $k(x_{k+1}, \cdot)$ is to consider a point x_{k+1} where the error bound is maximum, that is

$$x_{k+1} \in \arg\max_{x \in \mathcal{X}} \Lambda_k(x),$$

with

$$\Lambda_k(x) = \|k(x,\cdot) - \mathcal{I}_{W_k} k(x,\cdot)\|_H^2 = k(x,x) - k(x,\boldsymbol{x}) k(\boldsymbol{x},\boldsymbol{x})^{-1} k(\boldsymbol{x},x).$$

Interpolation in RKHS

A finite dimensional space V_m with basis $\varphi_1, \dots, \varphi_m$ defines a RKHS with kernel

$$k(x, y) = \boldsymbol{\varphi}(x)^T \boldsymbol{\varphi}(y), \quad \boldsymbol{\varphi}(x) := (\varphi_1(x), \dots, \varphi_m(x))$$

A sequential interpolation method consists in defining a sequence of points $(x_k)_{k \geq 1}$ and corresponding spaces $W_k = span\{k(x_1, \cdot), \dots, k(x_k, \cdot)\}$ such that

$$x_{k+1} = \arg\max_{x \in \mathcal{X}} \Lambda_k(x),$$

where

$$\Lambda_k(x) = \|\boldsymbol{\varphi}(x)\|_2^2 - \boldsymbol{\varphi}(x)^T \boldsymbol{\varphi}(\boldsymbol{x})(\boldsymbol{\varphi}(\boldsymbol{x})\boldsymbol{\varphi}(\boldsymbol{x})^T)^{-1} \boldsymbol{\varphi}(\boldsymbol{x})^T \boldsymbol{\varphi}(x)$$

with $\boldsymbol{x} = (x_1, \dots, x_k)$ and $\boldsymbol{\varphi}(\boldsymbol{x}) = (\varphi_i(x_j))_{1 \leq i \leq m, 1 \leq j \leq k}$.

In bayesian regression with gaussian processes (with noisy-free observations), the function $\Lambda_k(x)$ is the variance of the conditional gaussian process given observations at points $\boldsymbol{x} = (x_1, \dots, x_k)$.

Note that the obtained sequence of points only depends on the space V_m.

Letting $U_k = span\{\boldsymbol{\varphi}(x_1), \dots, \boldsymbol{\varphi}(x_k)\} \subset \mathbb{R}^m$, we note that

$$\Lambda_k(x) = \|\boldsymbol{\varphi}(x) - P_{U_{k-1}} \boldsymbol{\varphi}(x)\|_2^2$$

This is equivalent to the previously presented empirical interpolation based on feature map.

Least squares approximation

Consider the approximation of a function f in $V = L^2_\mu(\mathcal{X})$ equipped with the norm

$$\|f\|^2 = \int f(x)^2 d\mu(x).$$

We are given a m-dimensional space V_m in $L^2_\mu(\mathcal{X})$.

A weighted least-squares approximation $\hat{f}_m \in V_m$ is defined by minimizing

$$\frac{1}{n}\sum_{i=1}^{n} w(x_i)(f(x_i) - v(x_i))^2 := \|f - v\|_n^2$$

over $v \in V_m$, for some suitably chosen points $\boldsymbol{x} = (x_1, \ldots, x_n)$ and weight function w.

If x_i are samples from a distribution $\nu = w^{-1}\mu$, then

$$\mathbb{E}(\|\cdot\|_n^2) = \|\cdot\|^2$$

Least squares approximation

Given an L^2_μ-orthonormal basis $\varphi_1(x), ..., \varphi_m(x)$ of V_m,

$$\lambda_{min}(\boldsymbol{G})\|v\|^2 \leq \|v\|_n^2 \leq \lambda_{max}(\boldsymbol{G})\|v\|^2 \quad \forall v \in V_m,$$

where $\boldsymbol{G}$ is the empirical Gram matrix given by

$$\boxed{\boldsymbol{G} = \frac{1}{n}\sum_{i=1}^{n} w(x_i)\boldsymbol{\varphi}(x_i)\boldsymbol{\varphi}(x_i)^T}$$

with $\boldsymbol{\varphi}(x) = (\varphi_1(x), ..., \varphi_m(x))^T \in \mathbb{R}^m$.

The quality of least-squares projection is related to how much $\boldsymbol{G}$ deviates from the identity. Indeed

$$\begin{aligned}
\|f - \hat{f}_m\|^2 &\leq \|f - P_{V_m} f\|^2 + \|P_{V_m} f - \hat{f}_m\|^2 \\
&\leq \|f - P_{V_m} f\|^2 + \lambda_{min}(\boldsymbol{G})^{-1}\|P_{V_m} f - \hat{f}_m\|_n^2 \\
&\leq \|f - P_{V_m} f\|^2 + \lambda_{min}(\boldsymbol{G})^{-1}\|P_{V_m} f - f\|_n^2
\end{aligned}$$

where we have used the fact that $\hat{f}_m$ is the orthogonal projection of f onto V_m w.r.t. the semi-norm $\|\cdot\|_n$.

Least-squares approximation with i.i.d. sampling and conditioning

If the x_i are samples from $\nu = w^{-1}\mu$,

$$\mathbb{E}(\boldsymbol{G}) = \boldsymbol{I}$$

For i.i.d. samples, $\boldsymbol{G} := \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{A}(x_i)$ where the matrices $\boldsymbol{A}(x_i) := w(x_i)\boldsymbol{\varphi}(x_i)\boldsymbol{\varphi}(x_i)^T$ are i.i.d. and with spectral norm almost surely bounded by

$$K_w(V_m) = \sup_{x\in\mathcal{X}} w(x)\|\boldsymbol{\varphi}(x)\|_2^2.$$

From matrix Chernoff inequality [Tropp 2010, Cohen and Migliorati 2017], we know that

$$\mathbb{P}(\lambda_{max}(\boldsymbol{G}) > 1+\delta) \wedge \mathbb{P}(\lambda_{min}(\boldsymbol{G}) < 1-\delta) \leq m\exp(-\frac{n\delta^2}{2K_w(V_m)})$$

Therefore, provided

$$n \geq 2K_w(V_m)\delta^{-2}\log(m\eta^{-1})$$

it holds

$$\mathbb{P}(\lambda_{min}(\boldsymbol{G}) < 1-\delta) \leq \eta.$$

Classical least-squares approximation with i.i.d. sampling

For classical least-squares, $w = 1$ ($\nu = \mu$).

- For V_m piecewise constant functions on a uniform partition of $(0,1)$ and μ the uniform measure, $\boxed{K_{1,m} = m}$.
- For V_m trigonometric polynomials of degree $(m-1)/2$ on $(0, 2\pi)$ and μ the uniform measure, $\boxed{K_{1,m} = m}$.
- For polynomial spaces $V_m = \mathbb{P}_{m-1}$ and μ the uniform measure, $\boxed{K_{1,m} = m^2}$.
- For polynomial spaces $V_m = \mathbb{P}_{m-1}$ and μ the gaussian measure on $\mathbb{R}$, $\boxed{K_{1,m} = \infty}$.

Optimal measure for i.i.d. sampling

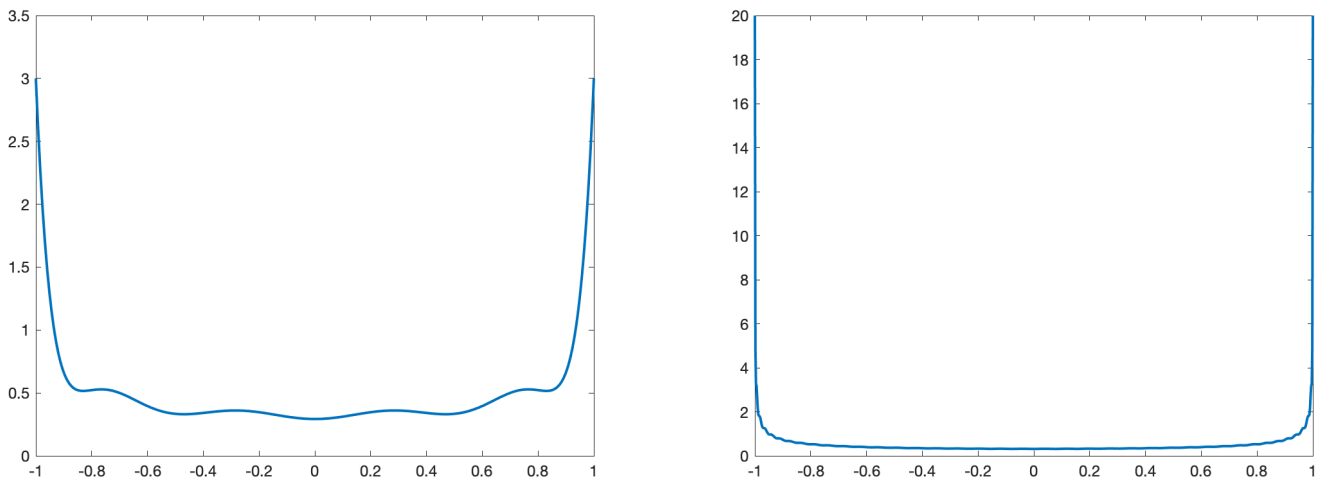
Optimal sampling measure (leverage score sampling) is given by

$$\nu_m = w_m^{-1}\mu \quad \text{with} \quad w_m(x)^{-1} = \frac{1}{m}\|\varphi(x)\|_2^2 = \frac{1}{m}\sum_{j=1}^{m}\varphi_j(x)^2 \quad \text{(Inverse Christoffel function)}$$

This gives an optimal constant $K_{w_m}(V_m) = m$.

Optimal weighted least squares with i.i.d. sampling

- For V_m piecewise constant functions on a uniform partition of $(0,1)$ and μ the uniform measure, $w_m(x) = 1$.
- For V_m trigonometric polynomials of degree $(m-1)/2$ on $(0, 2\pi)$ and μ the uniform measure, $w_m(x) = 1$.
- For polynomial spaces $V_m = \mathbb{P}_{m-1}$ and μ the uniform measure on $(-1,1)$

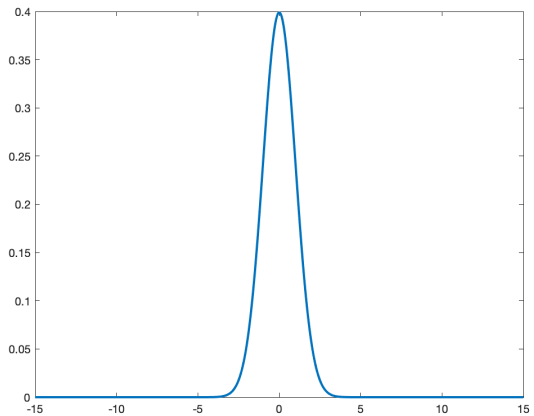


(a) $m = 6$ (b) $m = 40$

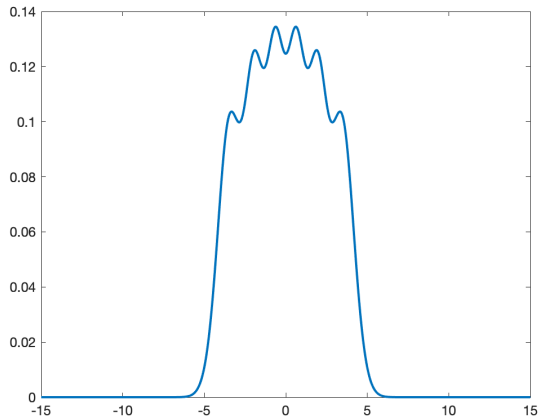
Figure: Polynomials and uniform measure: density of ν_m

Optimal weighted least squares with i.i.d. sampling

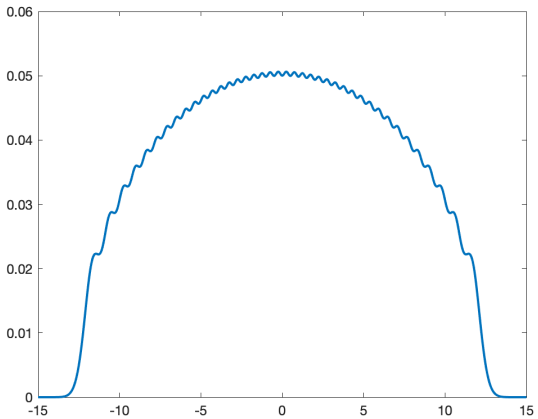
- For polynomial spaces $V_m = \mathbb{P}_{m-1}$ and μ the standard gaussian measure on $\mathbb{R}$



(a) $m = 0$



(b) $m = 6$



(c) $m = 40$

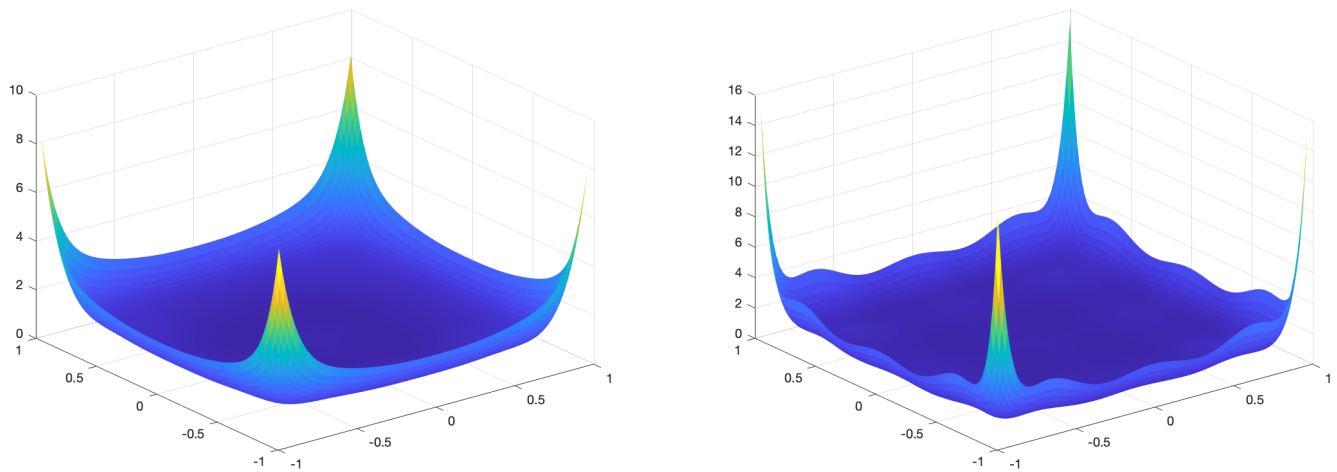
Figure: Polynomials and Gaussian measure: density of $\nu_m = w_m^{-1}\mu$

Optimal weighted least squares with i.i.d. sampling

- For *d*-variate polynomials,

$$V_m = \mathbb{P}_\Lambda := span\{x^\alpha = x_1^{\alpha_1} \dots x_d^{\alpha_d} : \nu \in \Lambda \subset \mathbb{N}^d\}$$

$\Lambda = \Lambda_{1,p} := \{\alpha : \|\alpha\|_1 \leq p\}$ corresponds to polynomials with total degree $\leq p$.
$\Lambda = \Lambda_{\infty,p} := \{\alpha : \|\alpha\|_\infty \leq p\}$ corresponds to polynomials with partial degree $\leq p$.



(a) $\Lambda_{1,4}$ (b) $\Lambda_{\infty,4}$

Figure: Polynomials and uniform measure on $[-1,1]^2$:
density w_m for polynomials with total (left) or partial (right) degree less than 4.

Sampling from the optimal measure (i.i.d. setting)

We have to sample from the optimal measure

$$d\nu_m = w_m^{-1} d\mu, \quad w_m(x)^{-1} = \frac{1}{m} \sum_{j=1}^{m} \varphi_j(x)^2$$

Standard sampling technique can be used: inverse transform, rejection, Markov Chain Monte-Carlo...

However, for general spaces V_m, sampling may be a non trivial task.

Sampling from the optimal measure: mixture sampling

We observe that ν_m is a mixture of measures

$$d\nu^{(j)}(x) = \varphi_j(x)^2 d\mu(x)$$

with equal weights $1/m$. We can first sample j uniformly at random in $\{1, \ldots, m\}$ and then sample from $\nu^{(j)}$.

Recycling samples for adaptive approximation

In adaptive approximation, we construct approximations from a sequence of spaces $(V_m)_{m \geq 1}$.

To each space V_m is associated a specific optimal sampling measure $\nu_m = w_m^{-1} \mu$.

When functions evaluations are costly, we would like to exploit samples generated at previous iterations.

Recycling samples for adaptive approximation: hierarchical spaces

Consider the adaptive approximation in a sequence of nested spaces

$$V_1 \subset \ldots \subset V_m \subset V_{m+1} \subset \ldots$$

Let $(\varphi_j)_{j\geq 1}$ be such that $V_m = span\{\varphi_1, \ldots, \varphi_m\}$. Then

$$V_{m+1} = V_m \oplus span\{\varphi_{m+1}\}$$

and the optimal sampling measure ν_{m+1} associated to V_{m+1} is such that

$$d\nu_{m+1}(x) = \frac{1}{m+1}\sum_{j=1}^{m+1}\varphi_j(x)^2 d\mu(x) = \frac{m}{m+1}d\nu_m(x) + \frac{1}{m+1}\varphi_{m+1}^2 d\mu(x)$$

that corresponds to a mixture between ν_m and $\varphi_{m+1}^2\mu$, with respective weights $\frac{m}{m+1}$ and $\frac{1}{m+1}$.

To sample the mixture, draw a Bernoulli variable $B(\frac{1}{m+1})$. If 1 is obtained, generate a new sample from $\varphi_{m+1}^2\mu$. If 0 is obtain, then either pick without replacement a sample from previously generated samples from ν_m, or generate a new sample from ν_m.

Different strategies can be found in [Arras et al 2019, Migliorati 2019].

Optimal weighted least-squares with conditioning

Assume that $\boldsymbol{x} = (x_1, \ldots, x_n)$ are drawn from $\nu_m^{\otimes n}$ conditioned to satisfy the event $S_\delta = \{\lambda_{min}(\boldsymbol{G}) > 1 - \delta\}$. This can be obtained by sampling $\boldsymbol{x}$ from $\nu_m^{\otimes n}$ until S_δ is satisfied (rejection).

Under the condition

$$n \geq m\delta^{-2} \log(2m\eta^{-1}) \tag{1}$$

we have

$$\mathbb{P}(S_\delta) \geq 1 - \eta$$

For $\eta < 1$, the random number N of samples from $\nu_m^{\otimes n}$ generated before acceptation follows a geometric distribution with parameter $\mathbb{P}(S_\delta)$, is almost surely finite, and with expectation $\mathbb{E}(N) = \mathbb{P}(S_\delta)^{-1} \leq (1 - \eta)^{-1}$.

The least-squares estimator satisfies

$$\begin{aligned}
\mathbb{E}(\|f - \hat{f}_m\|^2) &\leq \|f - f_m\|^2 + (1 - \delta)^{-1}\mathbb{E}(\|f - f_m\|_n^2) \\
&\leq \|f - f_m\|^2 + (1 - \delta)^{-1}(1 - \eta)^{-1}\mathbb{E}_{\boldsymbol{x} \sim \nu_m^{\otimes n}}(\|f - f_m\|_n^2) \\
&= (1 + (1 - \delta)^{-1}(1 - \eta)^{-1})\|f - f_m\|^2
\end{aligned}$$

Optimal weighted least-squares with conditioning

Therefore, we deduce a quasi-optimality in expectation

$$\mathbb{E}(\|f - \hat{f}_m\|^2)^{1/2} \leq C \inf_{v \in V_m} \|f - v\|,$$

with $C = (1 + (1-\delta)^{-1}(1-\eta)^{-1})^{1/2}$.

For a compact set K of functions in L^2_μ, using the previous result with an optimal subspace V_m of dimension m such that

$$\inf_{v \in V_m} \|f - v\| = d_m(K)_{L^2_\mu},$$

we deduce that for $n \geq cm\log(m)$, for some universal constant c, there exists a distribution over $\mathcal{X}^n$ and a linear recovery map A such that

$$\mathbb{E}(\|f - A(f(x_1), \ldots, f(x_n))\|^2)^{1/2} \leq C d_m(K)_{L^2_\mu}$$

which proves

$$\rho^{rand}_{cm\log(m)}(K)_{L^2_\mu} \leq C d_m(K)_{L^2_\mu}$$

Reducing the sampling complexity

The number of i.i.d. samples $n \sim \delta^{-2} m \log(m)$ may still be large compared to m, and a fundamental question is whether we can achieve stability with $n \sim m$.

One route is to rely on subsampling [Haberstich, Nouy and Perrin 2022] [Dolbeault and Cohen 2022] [Dolbeault, Krieg and Ullrich 2023] [Bartel, Schafer and T. Ullrich 2023], i.e. start with a large number of samples ensuring stability of the Gram matrix, and then select a (hopefully small) subset of samples preserving stability.

Another route is to introduce dependence between the samples to better control the spectrum of the Gram matrix. [Dolbeault and Chkifa 2024] introduce a sequential sampling algorithm inspired by subsampling algorithms, yielding quasi-optimality in expectation with minimal oversampling.

Introducing dependence by volume sampling

An indirect way to control the minimal eigenvalue of the empirical Gram matrix is to maximize its determinant $\det(\boldsymbol{G}(\boldsymbol{x}))$.

In a deterministic setting, this corresponds to *D*-optimal design of experiments and is related to maximum volume concept in linear algebra [Goreinov et al 2010, Fonarev et al 2016], or Fekete points in interpolation.

In a randomized setting, consider a sample $\boldsymbol{x} = (x_1, \ldots, x_m)$ of size m from

$$\boxed{d\gamma_m(\boldsymbol{x}) \propto \det(\boldsymbol{G}(\boldsymbol{x}))d\nu_m^{\otimes m}(\boldsymbol{x})}$$

that tends to promote high determinant of $\boldsymbol{G}(\boldsymbol{x})$ and high likelihood w.r.t. optimal i.i.d. sampling measure $\nu_m^{\otimes m}$.

Introducing dependence by volume sampling

For $V = L^2_\mu$, γ_m is the distribution of a **projection determinantal point process (DPP)** for V_m and reference measure μ [Lavancier et al 2015]

$$d\gamma_m(\boldsymbol{x}) = \frac{1}{m!}\det(\boldsymbol{\varphi}(\boldsymbol{x})^T\boldsymbol{\varphi}(\boldsymbol{x}))d\mu^{\otimes m}(\boldsymbol{x}), \quad \boldsymbol{\varphi}(\boldsymbol{x})^T = (\boldsymbol{\varphi}(x_1)\dots\boldsymbol{\varphi}(x_m)) \in \mathbb{R}^{m\times m}.$$

The density $\det(\boldsymbol{\varphi}(\boldsymbol{x})^T\boldsymbol{\varphi}(\boldsymbol{x}))$ introduces a repulsion between points (null density whenever $\boldsymbol{\varphi}(x_i) = \boldsymbol{\varphi}(x_j)$ for $i \neq j$), and promotes dissimilarity in the selected features $\boldsymbol{\varphi}(x_i)$.

The marginals are all equal to the optimal measure ν_m for i.i.d. sampling.

The conditional distribution of x_{k+1} given $(x_1, \dots, x_k)$ has an explicit expression

$$x_{k+1}|x_1, \dots, x_k \sim \frac{1}{m-k}\|\boldsymbol{\varphi}(x) - P_{U_k}\boldsymbol{\varphi}(x)\|_2^2 d\mu(x),$$

with $U_k = span\{\boldsymbol{\varphi}(x_1), \dots, \boldsymbol{\varphi}(x_k)\} \subset \mathbb{R}^m$. This allows to easily sample sequentially.

Note that it is equivalent to a randomized version of the empirical interpolation method based on feature maps (or adaptive gaussian process interpolation) where the point x_{k+1} is chosen to maximize $\|\boldsymbol{\varphi}(x) - P_{U_k}\boldsymbol{\varphi}(x)\|_2^2$.

How to improve stability ?

Stability can be ensured with higher probability

- by adding $n - m$ i.i.d. samples from μ (if μ is a probability measure), which corresponds to **volume sampling** [Poinas, Bardenet 2022]

$$\det(\boldsymbol{\varphi}(\boldsymbol{x})^T \boldsymbol{\varphi}(\boldsymbol{x})) d\mu^{\otimes n}(\boldsymbol{x})$$

  A natural approach for classical (non-weighted) least-squares, but bad performance compared to optimal i.i.d. sampling.

- by adding $n - m$ i.i.d. samples from ν_m, which corresponds to **volume-rescaled sampling** [Dereziński et al 2022]

$$d\gamma_n(\boldsymbol{x}) \propto \det(\boldsymbol{G}(\boldsymbol{x})) d\nu_m^{\otimes n}(\boldsymbol{x})$$

  It yields an unbiased estimate of the orthogonal projection, $\mathbb{E}(\hat{f}_m) = P_{V_m} f$, but the performance is similar to i.i.d. optimal sampling from $\nu_m^{\otimes n}$.

How to improve stability ?

- by using multiple samples from γ_m **(repeated DPP)** [Nouy and Michel 2024].

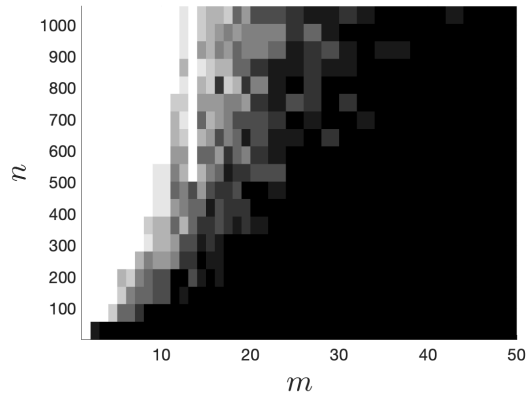
Theorem

Assume that $(x_1, \ldots, x_n)$ is drawn (by rejection) from $\gamma_m^{\otimes(n/m)}$ conditioned to the event $S_\delta = \{\lambda_{min}(\boldsymbol{G}) \geq 1 - \delta\}$. Then the weighted LS projection satisfies

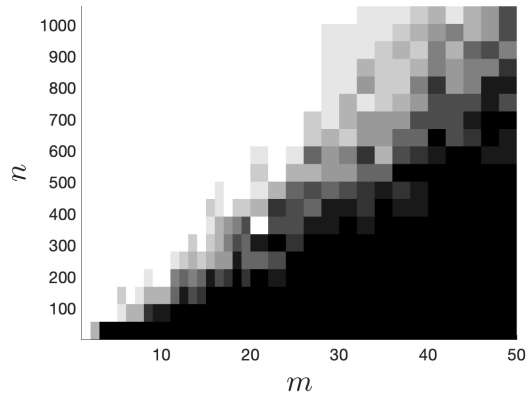
$$\mathbb{E}(\|f - \hat{f}_m\|^2) \leq (1 + \frac{m}{n}\mathbb{P}(S_\delta)^{-1}(1-\delta)^{-2}) \inf_{g \in V_m} \|f - g\|^2.$$

Similar theoretical guarantees as optimal i.i.d., but better concentration properties in practice.

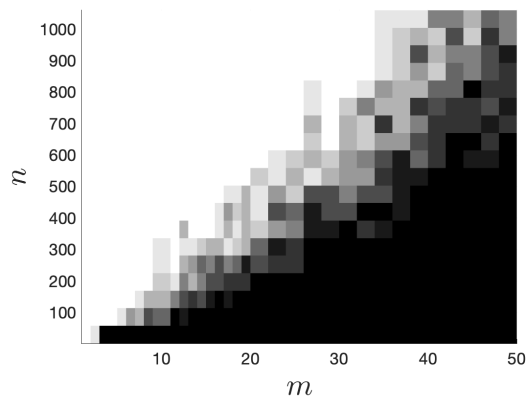
$\mathbb{P}(Sp(\boldsymbol{G}) \subset [1/2, 3/2])$ as a function of m and n



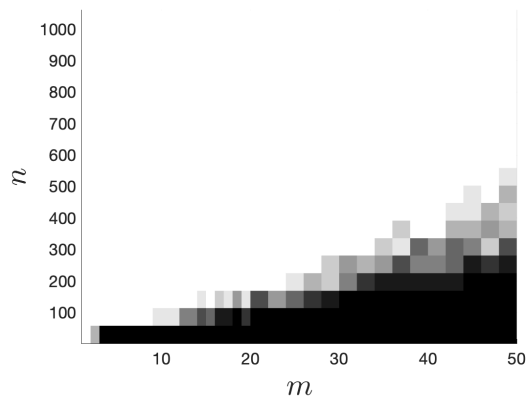
(a) i.i.d. μ (Classical)



(b) i.i.d. ν_m (Optimal i.i.d.)



(c) $\gamma_m + n - m$ i.i.d. ν_m (Volume-rescaled sampling)



(d) multiple γ_m (Repeated DPP)

Figure: $\mathbb{P}(Sp(\boldsymbol{G}) \subset [\frac{1}{2}, \frac{3}{2}])$ as a function of m and n, from 0 (black) to 1 (white). V_m is a polynomial space of degree $m-1$ and μ the uniform measure over $[-1, 1]$.

Other metrics, other information

These results can be extended to a Hilbert space V of functions equipped with a (semi-)norm

$$\|f\|_V^2 = \int_{\mathcal{X}} |L_x f|^2 d\mu(x)$$

where $L_x : \mathcal{H} \to \mathbb{R}^k$ is linear, using information $\ell_i(f) = L_{x_i} f$.

For example:

- $V = L^2(\mathcal{X}, \mu)$ for $L_x f = f(x)$
- $V = H^s(\mathcal{X}, \mu)$, $\mathcal{X} \subset \mathbb{R}^d$, with $L_x f = (D^\alpha f(x))_{\alpha \in \Lambda}$, $\Lambda = \{\alpha : |\alpha| \leq s\}$
- $V = H^s(\mathcal{X}, \mu)$ with $L_x f = \sum_\alpha a_\alpha(x) D^\alpha f(x)$ (differential operator)
- $V = H^s(\mathcal{Y}, \rho)$, $\mathcal{Y} \subset \mathbb{R}^d$, $\mathcal{X} = \mathcal{Y} \times \Lambda$, $\mu = \rho \otimes (\sum_{\alpha \in \Lambda} \delta_\alpha)$, $L_x f = D^\alpha f(y)$ for $x = (y, \alpha)$

Other metrics, other information

A weighted least-squares approximation $\hat{f}_m \in V_m$ is defined by minimizing over $v \in V_m$

$$\frac{1}{n}\sum_{i=1}^{n} w(x_i)^{-1}|L_{x_i}f - L_{x_i}v|^2 := \|f - v\|_n^2, \quad x_i \sim w\mu.$$

Given a V-orthonormal basis $\varphi_1, ..., \varphi_m$ of V_m,

$$\lambda_{min}(\boldsymbol{G})\|v\|_V^2 \leq \|v\|_n^2 \leq \lambda_{max}(\boldsymbol{G})\|v\|_V^2 \quad \forall v \in V_m,$$

where $\boldsymbol{G}$ is the empirical Gram matrix given by

$$\boxed{\boldsymbol{G} = \frac{1}{n}\sum_{i=1}^{n} w(x_i) L_{x_i}\boldsymbol{\varphi} L_{x_i}\boldsymbol{\varphi}^T}$$

with $L_x\boldsymbol{\varphi} = (L_x\varphi_1, \ldots, L_x\varphi_m)^T \in \mathbb{R}^{m\times \ell}$.

For i.i.d. samples, $\boldsymbol{G} := \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{A}(x_i)$ where the matrices $\boldsymbol{A}(x_i) := w(x_i)L_{x_i}\boldsymbol{\varphi}L_{x_i}\boldsymbol{\varphi}^T$ are i.i.d. and with spectral norm almost surely bounded by

$$K_w(V_m) = \sup_{x\in\mathcal{X}} w(x)\|L_x\boldsymbol{\varphi}\|_2^2.$$

Other metrics, other information

An optimal sampling measure (leverage score sampling for L^2_μ) is given by

$$\boxed{\nu_m = w_m^{-1}\mu \quad \text{with} \quad w_m(x)^{-1} = \frac{1}{c_m}\|L_x\varphi\|_2^2 \quad \text{(Inverse generalized Christoffel function)}}$$

This gives an optimal constant $K_{w_m}(V_m) = c_m \leq m$.

With conditioned sampling and $\mathcal{O}(m\log(m))$ samples, we prove quasi-optimality result in expectation in the V norm [Gruhlke, Nouy and Trunschke 2024]

$$\mathbb{E}(\|f - \hat{f}_m\|_V^2)^{1/2} \leq C \inf_{g\in V_m} \|f - g\|_V.$$

Volume sampling [Nouy and Michel 2024] can also be generalized to this setting.

Almost sure error bounds

We would like to obtain quasi-optimality guarantees almost surely. This requires further assumptions on the target function and a suitable correction of the weighted least-squares projection.

A weighted least-squares approximation satisfies

$$\|f - \hat{f}_m\|_V \leq \|f - g\|_V + \lambda_{min}(\boldsymbol{G})^{-1/2}\|f - g\|_n, \quad \forall g \in V_m$$

We require almost sure control of $\lambda_{min}(\boldsymbol{G})^{-1} \leq (1-\delta)^{-1}$ (by conditioning) and of the empirical norm $\|\cdot\|_n$.

Assuming the target function is in a subspace H such that for all $g \in H$,

$$\|g\|_V \leq C_H\|g\|_H \quad \text{(continuous embedding } H \hookrightarrow V\text{)}$$

and

$$\|g\|_n \leq C'_H\|g\|_H,$$

it holds almost surely

$$\|f - \hat{f}_m\|_V \leq (C_H + C'_H(1-\delta)^{-1/2}) \inf_{v \in V_m} \|f - v\|_H$$

Assume that there exists a positive density $h > 0$ such that

$$\operatorname*{ess\,sup}_{x \in \mathcal{X}} h(x)^{-1/2} |L_x g| \leq \|g\|_H, \quad \forall g \in H$$

For example

- $V = L^2_\mu(\mathcal{X})$, $H = L^\infty_\mu(\mathcal{X})$ and $h(x) = 1$. $C^2_H = \mu(\mathcal{X})$.
- H a RKHS continuously embedded in $V = L^2_\mu(\mathcal{X})$ with kernel k and $h(x) = k(x, x)$. $C^2_H = \int_{\mathcal{X}} k(x, x) d\mu(x)$.

Then by choosing for the density a mixture

$$w(x)^{-1} = \frac{1}{2} w_m(x)^{-1} + \frac{1}{2} h(x)$$

it holds

$$\|g\|_n \leq 2\|g\|_H \quad \text{and} \quad K_w(V_m) = \sup_{x \in \mathcal{X}} w(x) \|L_x \varphi\|_2^2 \leq 2 K_{w_m}(V_m) = 2c_m$$

Only a factor 2 is lost in the number of i.i.d. samples required to ensure $\lambda_{min}(\boldsymbol{G})^{-1} \leq (1 - \delta)^{-1}$ with controlled probability.

We can also generalize volume sampling and obtain similar guarantees [Nouy and Michel 2024]

Sampling numbers

Using subsampling techniques from [Cohen and Dolbeault 2021], we prove that for $H = L^\infty$ or H a RKHS associated with a trace class operator, there exists a set of $n \leq cm$ points and a linear algorithm such that for all $f \in H$, the produced approximation $\hat{f}_m = A(f(x_1), \dots, f(x_n))$ is such that

$$\|f - \hat{f}_m\|_V \leq C \inf_{g \in V_m} \|f - g\|_H$$

Consider a compact set $K \subset H$ and an optimal approximating subspace V_m in the sense that $\sup_{f \in K} E(f; V_m)_H = d_m(K)_H$. We then have proven that

$$\rho_{cm}(K)_{L^2} \leq C d_m(K)_H$$

For K the unit ball of a RKHS (with the trace class assumption), a refined analysis (see [Dolbeault, Krieg and Ullrich 2023]) yields

$$\rho_{cm}(K)_{L^2} \leq \sqrt{\frac{1}{m} \sum_{k \geq m} d_k(K)_{L^2}^2}$$

for some universal constant c, which is known as a sharp bound.

For a larger class of spaces including the space of bounded functions equipped with the supremum norm, it holds

$$\rho_{cm}(K)_{L^2} \leq \left(\frac{1}{m} \sum_{k \geq m} d_k(K)_{L^2}^p \right)^{1/p} \qquad \text{for any } 0 < p < 2$$

Outline

Nonlinear approximation: theory to practice gap

For a nonlinear manifold M described by m parameters, for obtaining an approximation $\hat{f}_m \in M$ with an error close to

$$\inf_{v \in M} \|f - v\|$$

the required number of samples n can be much higher than the number of parameters m.

- This is the theory to practice gap, proven for neural networks [Grohs and Voigtlander 2021] and tensor networks for i.i.d. samples [Eigel, Schneider and Trunschke, 2022].

- Quasi-optimality can be proven with i.i.d. sampling provided some condition $n \gtrsim K_w(M)$ [Trunschke 2022, Cardenas, Adcock, Dexter 2024], which yields an optimal sampling strategy (only depending on M), but with unreasonable sampling complexity when M is a highly nonlinear manifold.

  E.g. for sets M of low-rank tensors in a tensor space $U^{\otimes d}$, $K_w(M) = K_w(U^{\otimes d})$, that yields the condition $n \gtrsim \dim(U)^d$ (curse of dimensionality).

- More assumptions on functions are needed and algorithms and sampling should (in general) be adaptive.

Variational setting

Consider that f is solution of an optimization problem

$$\min_{v \in V} \mathcal{L}(v),$$

where $\mathcal{L}$ is some loss functional, e.g.

$$\mathcal{L}(v) = \frac{1}{2}\|f - v\|_V^2,$$

or others for different machine learning or scientific machine learning tasks.

Consider a differentiable manifold M in the Hilbert space V and assume we have access to evaluations of $\nabla\mathcal{L}(v)$ for a given $v \in M$.

A natural gradient descent

A natural gradient algorithm (in V) for solving

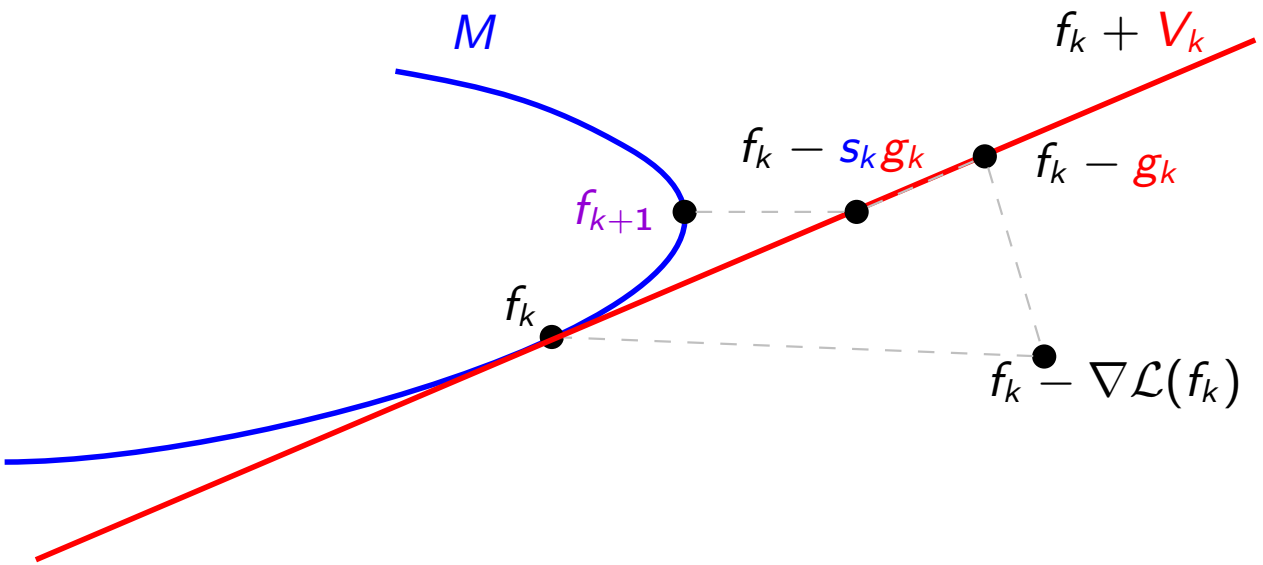
$$\min_{v \in M} \mathcal{L}(v)$$

constructs a sequence $(f_k)_{k \geq 0}$ by successive corrections in linear spaces V_k,

$$\boxed{f_{k+1} = R_k(f_k - s_k g_k)}$$

with

- $f_k + V_k$ is a local approximation of M
- g_k a projection of the gradient $\nabla \mathcal{L}(f_k)$ onto V_k
- s_k a step size
- R_k a retraction map with values in M



Optimal sampling for natural gradient descent[1]

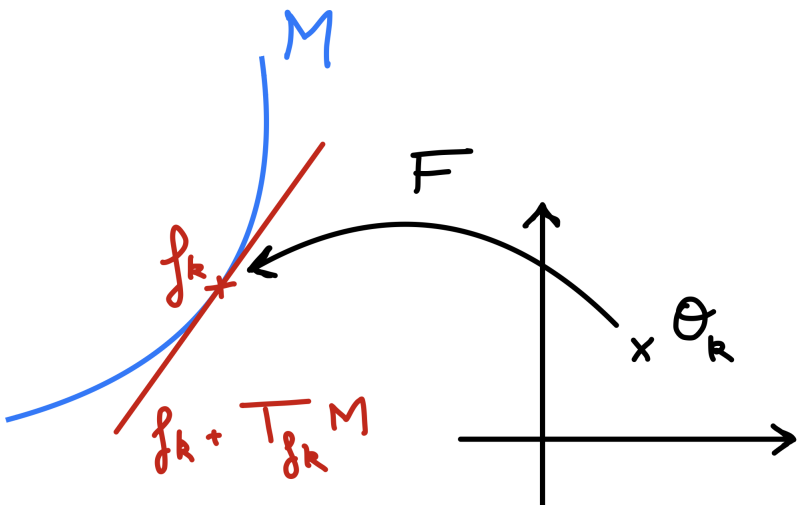
- g_k is defined as an empirical (quasi-)projection of the gradient onto V_k

$$g_k = \hat{P}_{V_k} \nabla \mathcal{L}(f_k)$$

using evaluations of $\nabla \mathcal{L}(f_k)$ at points drawn from an optimal sampling distribution for V_k.

- A natural choice for V_k is a linearization of $M = \{F(\theta) : \theta \in \mathbb{R}^m\}$ at $f_k = F(\theta_k)$,

$$T_{f_k} M = \text{span}\{\psi := \nabla_\theta F(\theta_k)\}$$



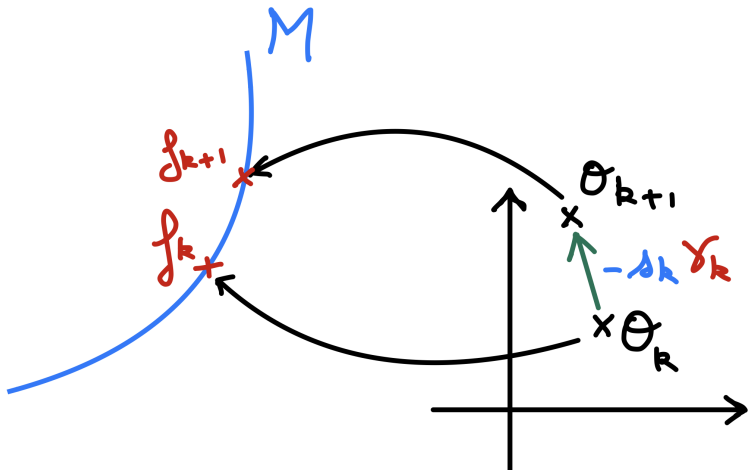
or a subspace of $T_{f_k} M$.

[1] R. Gruhlke, A. Nouy and P. Trunschke. Optimal sampling for stochastic and natural gradient descent: arXiv:2402.03113.

Optimal sampling for natural gradient descent

- A natural retraction is

$$f_{k+1} = R_k(f_k - s_k g_k) = F(\theta_{k+1}) \quad \text{with} \quad \theta_{k+1} = \theta_k - s_k \gamma_k \quad \text{and} \quad g_k(x) = \psi(x)^T \gamma_k.$$



- With $\mathcal{L}(v) = \int \ell(v(x); x) d\mu(x)$ estimated by $\mathcal{L}_n(v) = \frac{1}{n} \sum_{i=1}^n \ell(v(x_i); x_i)$, taking

$$\gamma_k = \nabla_\theta(\mathcal{L}_n(F(\theta_k))) = (\psi, \nabla\mathcal{L}(f_k))_n$$

corresponds to classical batch stochastic gradient descent (SGD), where g_k is a quasi-projection on V_k. It can be very far from the orthogonal projection of $\nabla\mathcal{L}(f_k)$.

- Using empirical projection yields a preconditioned SGD

$$\gamma_k = \boldsymbol{G}^{-1}(\psi, \nabla\mathcal{L}(f_k))_n \quad (\boldsymbol{G}\text{: empirical Gram of } \psi)$$

Application to neural networks

Consider the approximation of $f(x) = (1 + \prod_{i=1}^{d} x_i)^{-1}$, $x \in \mathcal{X} = [0,1]^d$ with $M = \{c^T \sigma(Ax + b) : c, b \in \mathbb{R}^s, A \in \mathbb{R}^{s \times d}\}$ (shallow neural network with softplus activation function $\sigma(t) = \log(1 + \exp(-x)))$.

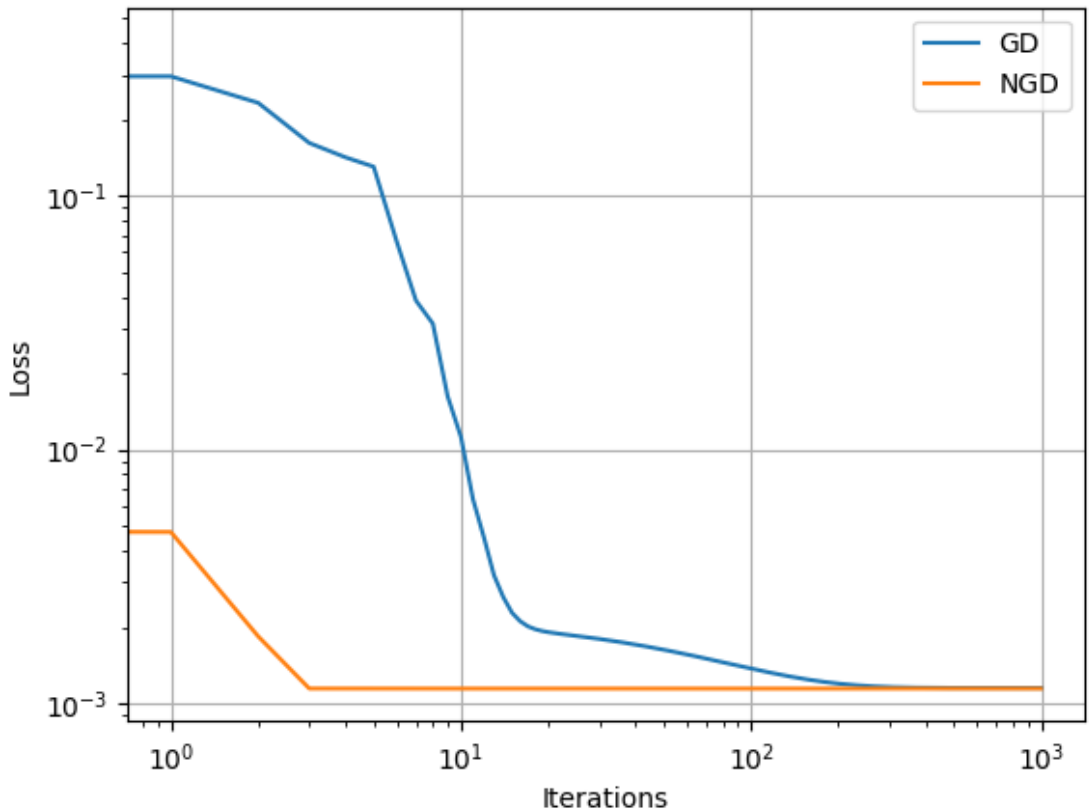


Figure: $d = 3$, $s = 6$. Comparison of GD and NGD given data.

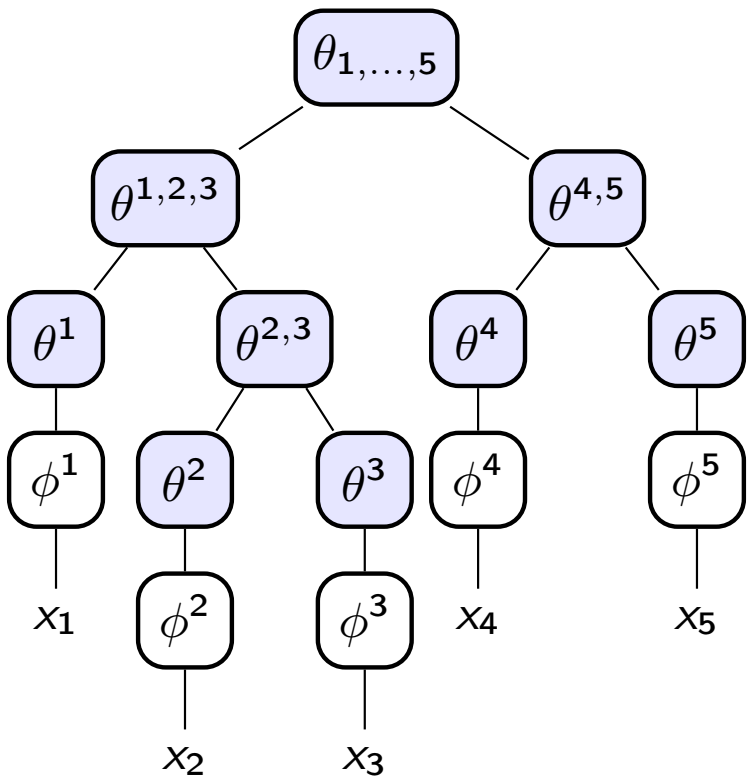
Application to tree tensor networks

Tree tensor networks form a prominent class of approximation tools for the approximation of multivariate functions $f(x_1, \ldots, x_d)$. This includes Tensor Train format [Oseledets & Tyrtyshnikov 2009], Hierarchical Tucker format [Hackbusch & Kuhn 2009].

They have a high approximation power (optimal rates for a large class of smoothness classes).

They admits a multilinear parametrization in terms of a collection of low-order tensors θ_α:

$$M = \{F(\theta_1, \ldots, \theta_L) : \theta_1 \in \mathbb{R}^{I_1}, \ldots, \theta_L \in \mathbb{R}^{I_L}\}, \quad F \text{ multilinear.}$$



Application to tree tensor networks

M is a differentiable manifold[2] with tangent space

$$T_{F(\theta)}M = span\{\nabla_{\theta_1} F(\theta)\} + \ldots + span\{\nabla_{\theta_L} F(\theta)\}$$

Controlled retraction using higher order singular value decomposition.

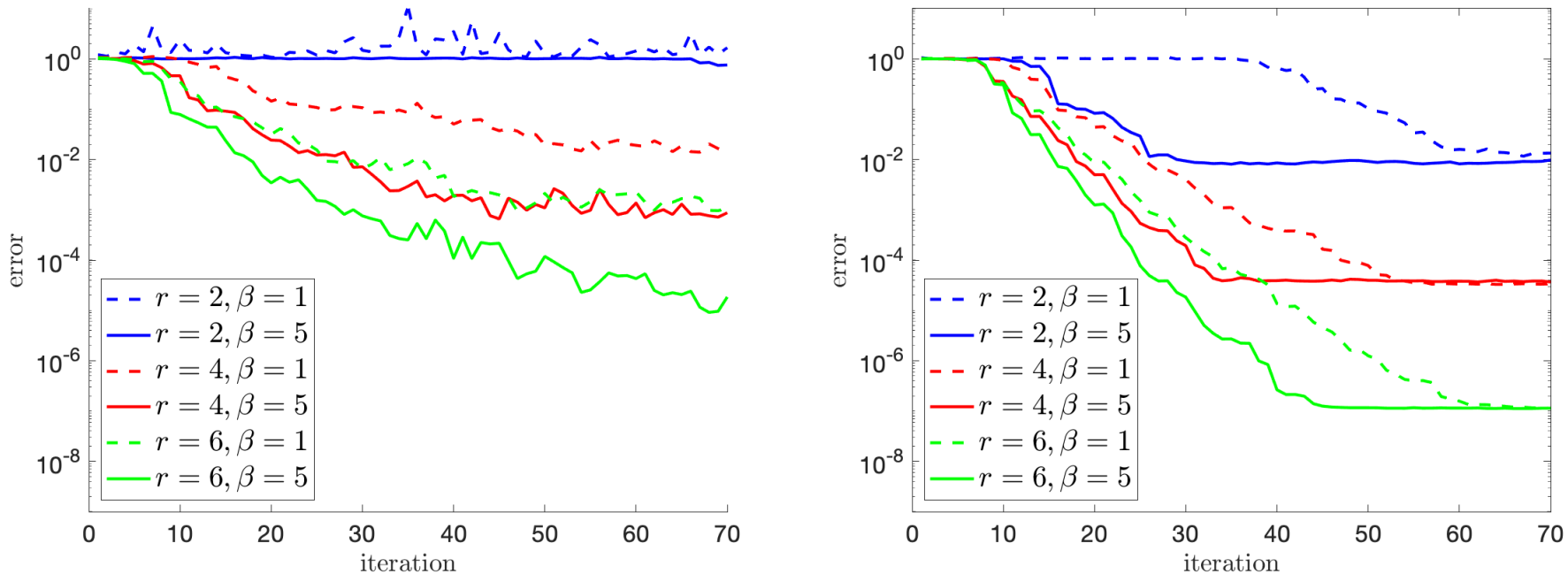
Choosing V_k as $span\{\nabla_{\theta_i} F(\theta)\}$ corresponds to coordinate descent (alternating minimization). No retraction is needed.

Using classical linear algebra, we obtain optimal sampling density in a format amenable for sequential sampling in high dimension.

[2] A. Falcó, W. Hackbusch, and A. Nouy. Geometry of tree-based tensor formats in tensor banach spaces. *Annali di Matematica Pura ed Applicata (1923 -)*, 2023.

Application to tree tensor networks

Approximation of function $f(x) = (1 + \sum_{i=1}^{d} x_i)^{-1}$ on $[0,1]^d$ $(d = 5)$ using tensor train format. Use of alternating minimization with step size $s = 1$.



(a) Classical i.i.d. sampling (no conditioning)

(b) Optimal i.i.d. sampling (conditioning)

Figure: Error versus iteration for different ranks and different oversampling factors β, where $n = \beta 4d \log(4d)$, $d = \dim(V_k)$.

Convergence analysis

Convergence results under assumptions on manifolds (smoothness and convexity) and assumptions on empirical projections, satisfied by empirical projections using i.i.d. samples from optimal distribution or (repeated) volume sampling.

Results are obtained for general loss functionals $\mathcal{L}$ under standard smoothness assumptions on $\mathcal{L}$.

See [Gruhlke, Nouy and Trunschke 2024].

Conclusions

- Theory of sampling well advanced for linear approximation and rather general type of information, for a given class V_m.
- Some challenging questions in adaptive settings (recycling information).
- Theory of optimal sampling for nonlinear approximation is very limited.
- Natural gradient methods for nonlinear approximation allow to use optimal sampling for linear approximation, with some convergence guarantees.

  Applies to a large class of risk functionals and metrics... towards physics informed optimal sampling and other machine learning tasks.
- Sampling can be efficiently implemented for some model classes (tree tensor networks and shallow networks in L^2 setting). Still some computational challenges for general nonlinear classes (deep networks) and risk functionals.

References I

Sampling and linear approximation

Y. Maday, N. C. Nguyen, A. T. Patera, and G. S. H. Pau.

A general multipurpose interpolation procedure: the magic points.

Communications On Pure and Applied Analysis, 8(1):383–404, 2009.

A. Cohen and G. Migliorati.

Optimal weighted least-squares methods.

SMAI Journal of Computational Mathematics, 3:181–203, 2017.

A. Cohen and M. Dolbeault.

Optimal pointwise sampling for l^2 approximation, 2021.

M. Dolbeault and A. Cohen.

Optimal pointwise sampling for $L2$ approximation.

Journal of Complexity, 68:101602, 2022.

M. Dolbeault, D. Krieg, and M. Ullrich.

A sharp upper bound for sampling numbers in l2.

Applied and Computational Harmonic Analysis, 63:113–134, 2023.

M. Dolbeault and M. A. Chkifa.

Randomized Least-Squares with Minimal Oversampling and Interpolation in General Spaces.

SIAM J. Numer. Anal., July 2024.

References II

B. Arras, M. Bachmayr, and A. Cohen.

Sequential sampling for optimal weighted least squares approximations in hierarchical spaces.

SIAM Journal on Mathematics of Data Science, 1(1):189–207, 2019.

C. Haberstich, A. Nouy, and G. Perrin.

Boosted optimal weighted least-squares.

Mathematics of Computation, 91(335):1281–1315, 2022.

G. Migliorati.

Adaptive approximation by optimal weighted least-squares methods.

SIAM Journal on Numerical Analysis, 57(5):2217–2245, 2019.

C. Haberstich.

Adaptive approximation of high-dimensional functions with tree tensor networks for Uncertainty Quantification.

Theses, École centrale de Nantes, Dec. 2020.

A. W. Marcus, D. A. Spielman, and N. Srivastava.

Interlacing families ii: Mixed characteristic polynomials and the kadison—singer problem.

Annals of Mathematics, pages 327–350, 2015.

References III

S. Nitzan, A. Olevskii, and A. Olevskii.

Exponential frames on unbounded sets.

Proceedings of the American Mathematical Society, 144(1):109–118, 2016.

F. Bartel, M. Schäfer, and T. Ullrich.

Constructive subsampling of finite frames with applications in optimal function recovery.

Applied and Computational Harmonic Analysis, 65:209–248, 2023.

V. Temlyakov.

On optimal recovery in L2.

Journal of Complexity, 65:101545, 2021.

N. Nagel, M. Schäfer, and T. Ullrich.

A new upper bound for sampling numbers.

Foundations of Computational Mathematics, pages 1–24, 2021.

A. Nouy and B. Michel.

Weighted least-squares approximation with determinantal point processes and generalized volume sampling.

SMAI Journal of Computational Mathematics, arXiv e-prints arXiv:2312.14057, 2024.

References IV

M. Dereziński, M. K. Warmuth, and D. Hsu.

Unbiased estimators for random design regression.

The Journal of Machine Learning Research, 23(1):7539–7584, 2022.

A. Poinas and R. Bardenet.

On proportional volume sampling for experimental design in general spaces.

Statistics and Computing, 33(1):29, 2022.

A. Belhadji, R. Bardenet, and P. Chainais.

Kernel interpolation with continuous volume sampling.

In *International Conference on Machine Learning*, pages 725–735. PMLR, Nov. 2020.

P. Trunschke.

Convergence bounds for nonlinear least squares for tensor recovery.

arXiv preprint arXiv:2208.10954, 2022.

P. Trunschke and A. Nouy.

Optimal sampling for least squares approximation with general dictionaries.

arXiv e-prints arXiv:2407.07814, 2024.

P. Trunschke and A. Nouy.

Almost-sure quasi-optimal approximation in reproducing kernel hilbert spaces.

arXiv e-prints arXiv:2407.06674, 2024.

References V

M. Dolbeault and M. A. Chkifa.

Randomized Least-Squares with Minimal Oversampling and Interpolation in General Spaces.

SIAM J. Numer. Anal., July 2024.

P. Grohs and F. Voigtländer.

Proof of the theory-to-practice gap in deep learning via sampling complexity bounds for neural network approximation spaces.

CoRR, abs/2104.02746, 2021.

F. Lavancier, J. Møller, and E. Rubak.

Determinantal point process models and statistical inference.

Journal of the Royal Statistical Society. Series B (Statistical Methodology), 77(4):853–877, 2015.

J. A. Tropp.

User-friendly tail bounds for sums of random matrices.

Foundations of computational mathematics, 12(4):389–434, 2012.

Learning with tensor networks

B. Michel and A. Nouy.

Learning with tree tensor networks: complexity estimates and model selection.

arXiv e-prints, page arXiv:2007.01165, July 2020.

References VI

E. M. Stoudenmire and D. J. Schwab.

Supervised learning with quantum-inspired tensor networks, 2017.

E. Grelier, A. Nouy, M. Chevreuil.

Learning with tree-based tensor formats.

Arxiv eprints, Nov. 2018.

E. Grelier, A. Nouy, and R. Lebrun.

Learning high-dimensional probability distributions using tree tensor networks.

arXiv preprint arXiv:1912.07913, 2019.

A. Nouy.

Higher-order principal component analysis for the approximation of tensors in tree-based low-rank formats.

Numerische Mathematik, 141(3):743–789, Mar 2019.

C. Haberstich, A. Nouy, and G. Perrin.

Active learning of tree tensor networks using optimal least-squares.

arXiv preprint arXiv:2104.13436, 2021.

I. Oseledets and E. Tyrtyshnikov.

TT-cross approximation for multidimensional arrays.

Linear Algebra And Its Applications, 432(1):70–88, JAN 1 2010.

References VII

L. Grasedyck and S. Krämer.

Stable als approximation in the tt-format for rank-adaptive tensor completion.
Numerische Mathematik, 143(4):855–904, 2019.

Software

Nouy Anthony, Grelier Erwan and Giraldi Loic. (2020, February 7). ApproximationToolbox. Zenodo.
http://doi.org/10.5281/zenodo.3653970

Anthony Nouy, & Erwan Grelier. (2020, June 15). anthony-nouy/tensap. Zenodo.
http://doi.org/10.5281/zenodo.3894378

Sampling for nonlinear approximation

M. Eigel, R. Schneider, and P. Trunschke.

Convergence bounds for empirical nonlinear least-squares.
ESAIM: Mathematical Modelling and Numerical Analysis, 56(1):79–104, 2022.

R. Gruhlke, A. Nouy, and P. Trunschke.

Optimal sampling for stochastic and natural gradient descent.
arXiv e-prints arXiv:2402.03113, 2024.

J. M. Cardenas, B. Adcock, and N. Dexter.

Cs4ml: A general framework for active learning with arbitrary data based on christoffel functions.
Advances in Neural Information Processing Systems, 36, 2024.

Outline

Outline

Empirical interpolation

Given a space V_m with basis $\varphi_1, \ldots, \varphi_m$, a general greedy algorithm has been proposed in [1] to construct interpolation points, called magic points.

The idea is to construct a good sequence of spaces $W_k = span\{\psi_1, \ldots, \psi_k\}$ for the approximation of the discrete set $\{\varphi_i : 1 \leq i \leq m\}$ in $(X, \|\cdot\|_\infty)$, and associated interpolation points.

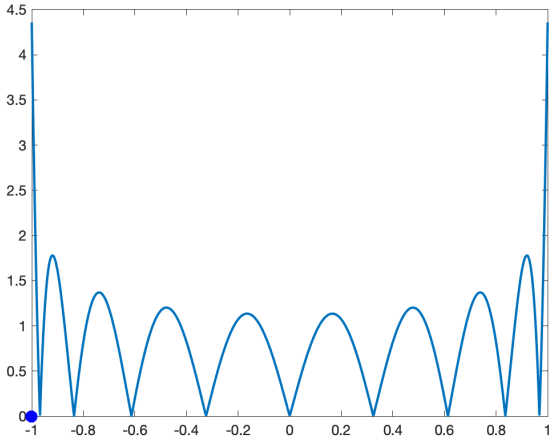
Starting from $V_0 = \{0\}$, we define

$$i_k \in \arg\max_{1 \leq i \leq m} \|\varphi_i - \mathcal{I}_{W_{k-1}} \varphi_i\|_\infty, \quad \psi_k = \varphi_{i_k} - \mathcal{I}_{W_{k-1}} \varphi_{i_k}$$

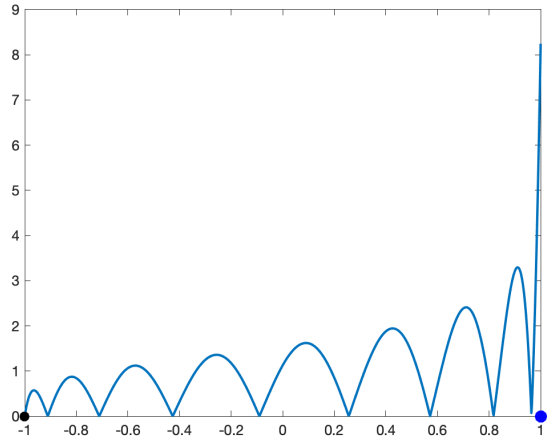
where $\mathcal{I}_{W_{k-1}}$ is the interpolation onto W_{k-1} using points $(x_1, \ldots x_{k-1})$, and define

$$x_k \in \arg\max_{x \in \mathcal{X}} |\psi_k(x)|.$$

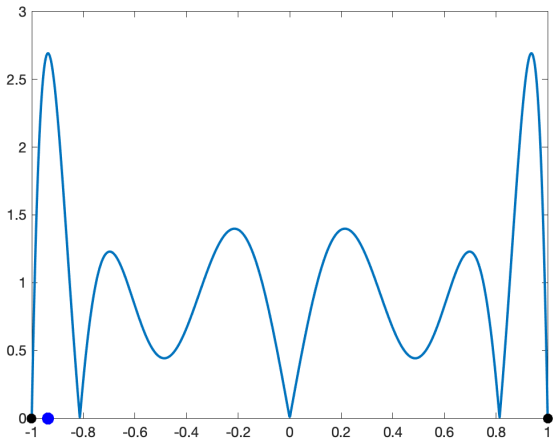
Empirical interpolation



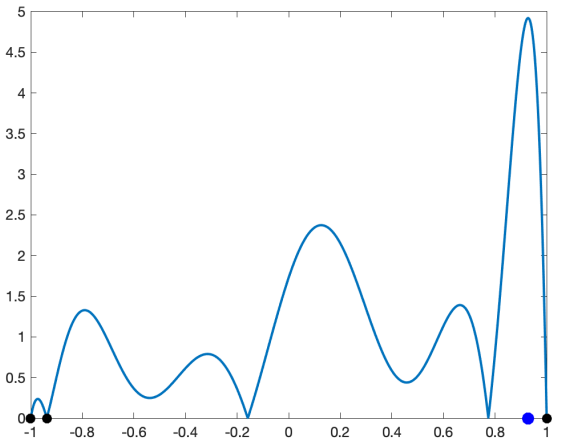
(a) $k = 1$



(b) $k = 2$



(c) $k = 3$



(d) $k = 4$

Figure: Polynomial space $V_m = \mathbb{P}_9$ on $[-1, 1]$. Function $|\psi_k(x)|$ and corresponding interpolation point $x_k = \arg\max_x |\psi_k(x)|$

Empirical interpolation

In the context of adaptive approximation in a sequence of spaces $V_1 \subset \ldots \subset V_m \subset \ldots$, and in order to recycle interpolation points, we modify the algorithm by simply taking $W_k = V_k$.

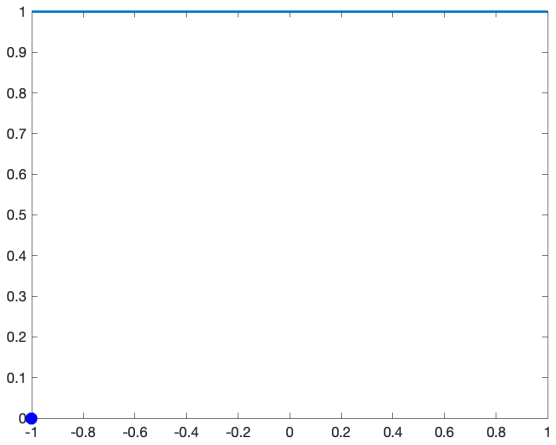
Letting $V_0 = \{0\}$, we define

$$\psi_k = \varphi_k - \mathcal{I}_{V_{k-1}} \varphi_k$$

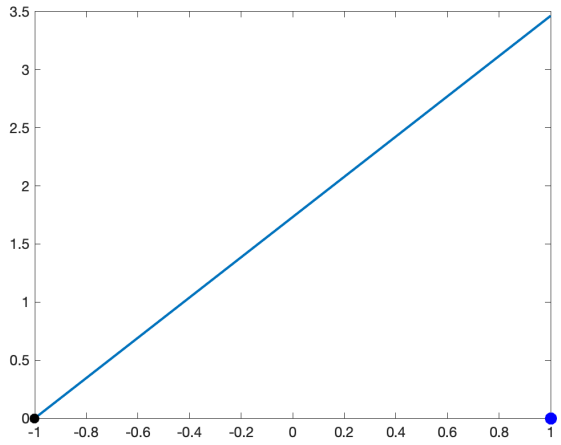
where $\mathcal{I}_{V_{k-1}}$ is the interpolation onto V_{k-1} using points $(x_1, \ldots x_{k-1})$, and define

$$x_k \in \arg\max_{x \in \mathcal{X}} |\psi_k(x)|.$$

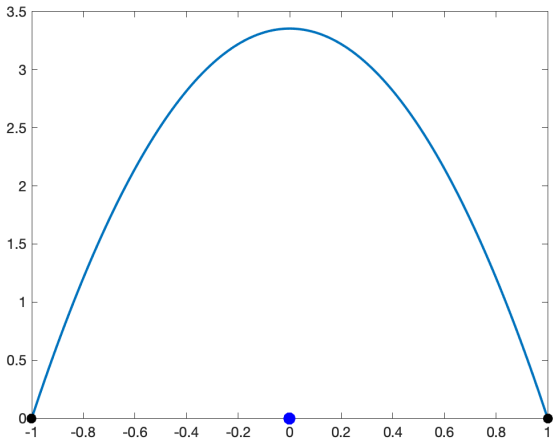
Empirical interpolation — adaptive setting



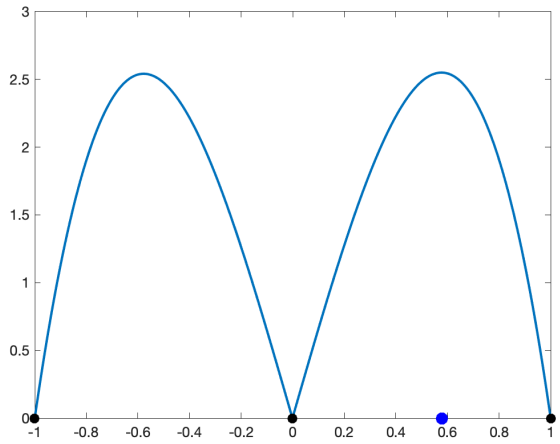
(a) $k = 1$



(b) $k = 2$



(c) $k = 3$



(d) $k = 4$

Figure: Polynomial space $V_m = \mathbb{P}_9$ on $[-1, 1]$. Function $|\psi_k(x)|$ and corresponding interpolation point $x_k = \arg\max_x |\psi_k(x)|$

Outline

Optimal design of experiments

Consider the model

$$Y = f(X) + \epsilon$$

where $X \sim \mu$ and $\epsilon \sim \mathcal{N}(0, \lambda)$ is independent of X, that corresponds to noisy evaluations of a function f.

For given points $\boldsymbol{x} = (x_1, \ldots, x_n)$ we have access to $\boldsymbol{y} = (y_1, \ldots, y_n)$ such that

$$y_i = f(x_i) + \epsilon_i$$

with $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n) \sim \mathcal{N}(0, \Lambda)$ independent of $\boldsymbol{x}$.

A weighted least-squares estimate $\hat{f}_m$ is then obtained by solving

$$\min_{v \in V_m} \frac{1}{n} \sum_{i=1}^{n} w_i (v(x_i) - y_i)^2$$

Letting $\Phi := \Phi(\boldsymbol{x}) = (\varphi_j(x_i))_{1 \leq i \leq n, 1 \leq j \leq m}$ (the design matrix) and $\boldsymbol{W} = diag(\boldsymbol{w})$ the weight matrix, we have

$$\hat{f}_m(x) = \boldsymbol{\varphi}(x)^T \hat{\boldsymbol{a}}, \quad \hat{\boldsymbol{a}} = \boldsymbol{G}^{-1} \Phi^T \boldsymbol{W} \boldsymbol{y}$$

with

$$\boldsymbol{G} := \boldsymbol{G}(\boldsymbol{x}, \boldsymbol{w}) = \Phi^T \boldsymbol{W} \Phi$$

Optimal design of experiments

For fixed $\boldsymbol{x}$, the covariance of $\hat{\boldsymbol{a}}$ is

$$Cov(\hat{\boldsymbol{a}}) = (\Phi^T \boldsymbol{W} \Phi)^{-1} \Phi^T \boldsymbol{W} \Lambda \boldsymbol{W} \Phi (\Phi^T \boldsymbol{W} \Phi)^{-1}$$

For $\Lambda = \lambda \boldsymbol{W}^{-1}$, we obtain

$$Cov(\hat{\boldsymbol{a}}) = \lambda \boldsymbol{G}^{-1}$$

and the variance of the prediction $\hat{f}_m(x)$ at some point x is

$$\mathbb{V}(\hat{f}_m(x)) = \lambda \boldsymbol{\varphi}(x)^T \boldsymbol{G}^{-1} \boldsymbol{\varphi}(x)$$

In order to minimize the variance for any $x \in \mathcal{X}$, that is for any $\boldsymbol{\varphi}(x) \in \mathbb{R}^m$, we would like to minimize $\boldsymbol{G}^{-1}$ over $\boldsymbol{x} \in \mathcal{X}^n$ and $\boldsymbol{w} \in \mathbb{R}_+^n$, in the sense of the Loewner order, over the space S_m^+ of symmetric positive semi-definite matrices. However, a global optimum does not necessarily exist since Loewner order is only a partial order.

Optimal design of experiments

A common approach is to consider as a proxy the minimization of a decreasing convex function $h : S_m^+ \to \mathbb{R}$, i.e. such that

$$h(\boldsymbol{A}) \leq h(\boldsymbol{B}) \quad \text{for} \quad \boldsymbol{A} \succcurlyeq \boldsymbol{B},$$

and solve

$$\min_{\boldsymbol{x}, \boldsymbol{w}} h(\boldsymbol{G}(\boldsymbol{x}, \boldsymbol{w}))$$

- E-optimal design corresponds $h(\boldsymbol{G}) = \lambda_{max}(\boldsymbol{G}^{-1}) = \lambda_{min}(\boldsymbol{G})^{-1}$
- A-optimal design corresponds to $h(\boldsymbol{G}) = Tr(\boldsymbol{G}^{-1})$
- D-optimal design corresponds to $h(\boldsymbol{G}) = det(\boldsymbol{G}^{-1}) = det(\boldsymbol{G})^{-1}$
- c-optimal design correspond to $h(\boldsymbol{G}) = \boldsymbol{c}^T \boldsymbol{G}^{-1} \boldsymbol{c}$ for some vector $\boldsymbol{c} \in \mathbb{R}^m$.

Outline

Convergence analysis

We make the following asumptions

- The empirical (quasi-)projection $\hat{P}_U$ onto a d-dimensional linear space U satisfies

$$(P_U g, \mathbb{E}(\hat{P}_U^n g - P_U g)) \geq -c_b \|P_U g\| \|(id - P_U)g\| \qquad \text{(bias)},$$
$$\mathbb{E}(\|\hat{P}_U^n g\|^2) \leq c_v \|g\|^2 \qquad \text{(variance)}$$

  where $c_b = c_b(n) \to 0$ as $n \to \infty$.

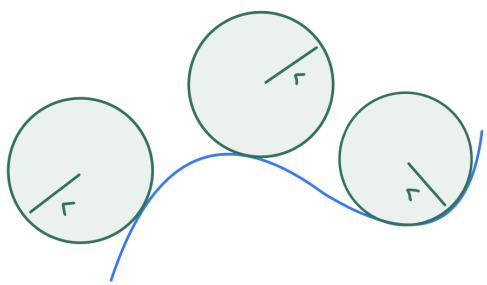
  Satisfied by (unbiased) quasi-projection or least-squares projections using i.i.d. samples from optimal distribution or (repeated) determinantal point processes. Requires a number of samples $n \lesssim d \log(d)$.

- The retraction map R_k at f_k satisfies

$$\mathcal{L}(R_k(f_k + g)) \leq \mathcal{L}(f_k + g) + \frac{C_R}{2}\|g\|^2 + \beta_k$$

  with some prescribed sequence $\beta_k = o(s_k)$.

  Requires an assumption on the reach (or curvature) of the manifold and adaptation of the step size.



Convergence analysis

With $(\mathcal{F}_k)_{k\geq 1}$ the filtration associated with the samples generated until step k, it holds

$$\mathbb{E}(\mathcal{L}(f_{k+1})|\mathcal{F}_k) \leq \mathcal{L}(f_k) - \gamma_k s_k \|P_{V_k}\nabla\mathcal{L}(f_k)\| + \frac{1+C_R}{2} c_v s_k^2 \|\nabla\mathcal{L}(f_k)\|^2 + \beta_k$$

where

$$\gamma_k = 1 - c_b \frac{\|(id - P_{V_k})\nabla\mathcal{L}(f_k)\|}{\|P_{V_k}\nabla\mathcal{L}(f_k)\|}$$

- For unbiased projections ($c_b = 0$) and step size s_k sufficiently small (deterministic)

  $$\mathbb{E}(\mathcal{L}(f_{k+1})|\mathcal{F}_k) \leq \mathcal{L}(f_k)$$

  We even obtain almost sure convergence using martingale theory ([Robbins and Siegmund 1971]), with algebraic rates between $\mathcal{O}(k^{-1})$ (GD) and $\mathcal{O}(k^{-1/2})$ (SGD).

  In favorable cases (recovery setting) and assuming strong Polyak-Lojasiewicz condition on manifold, we even get the exponential rate of GD, unlike SGD.
- For biased projections ($c_b > 0$), possible decay with sufficiently small step size only if $\gamma_k > 0$. Condition depending on the capacity of V_k to approximate the current gradient $\nabla\mathcal{L}(f_k)$. Feasible with sufficiently small c_b (large n).

  We prove a convergence towards a neighborhood of a stationary point.