Optimal sampling for approximation

Anthony Nouy

Centrale Nantes, Nantes Université,
Laboratoire de Mathématiques Jean Leray

Joint works with
Robert Gruhlke, Bertrand Michel, Philipp Trunschke

Approximation

We consider the approximation of a function f of a normed space V by an element of a subset 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 $V_m = \{F(\theta) : \theta \in \mathbb{R}^m\}$ which approximates K, obtained by manifold approximation (or model order reduction) 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 $L_1(f), \dots, L_n(f)$ and returning

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

where R is a reconstruction map with values in V_m.

Different types of information (context dependent)

- pointwise evaluations of the function

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

- pointwise evaluations of the function and its derivatives

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

- linear forms

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

  with B some operator (e.g. for solving $Bf = g$)

Approximation from limited information

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

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

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.

This requires to adapt the information to V_m and the target function class (active learning setting).

Outline

Least squares approximation

Consider the approximation of a function f from a Hilbert space V equipped with a norm

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

where $L_x : V \to \mathbb{R}^\ell$ is a linear map, e.g. $V = L^2_\mu(\mathcal{X})$ for $L_x f = f(x)$, $V = H^1_\mu(\mathcal{X})$ for $L_x f = \begin{pmatrix} f(x) \\ \nabla f(x) \end{pmatrix} \ldots$

Assume we can evaluate $L_x f$ for any $x \in \mathcal{X}$.

We are given a m-dimensional subspace V_m in V.

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

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

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 a V-orthonormal basis $\varphi_1, ..., \varphi_m$ 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

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

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

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

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

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

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 $\boldsymbol{A}(x_i) := w(x_i) L_{x_i}\boldsymbol{\varphi} L_{x_i}\boldsymbol{\varphi}^T$ are i.i.d. and with spectral norm almost surely bounded by

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

From matrix Chernoff inequality [**?**, **?**], we know that

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

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

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

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

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

Theorem ([Cohen and Migliorati 2017][Haberstich, N., Perrin 2022] [Gruhlke, N. and Trunschke 2024])

Assume that $(x_1, \dots, x_n)$ *is drawn (by rejection) from* $\nu_m^{\otimes n}$ *conditioned to the event*

$$S_\delta = \{\lambda_{min}(\boldsymbol{G}) \geq 1 - \delta\}, \quad 0 < \delta < 1,$$

and

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

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

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

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 [**?**] [**?**] [**?**].

Another route is to introduce dependence between the samples to better control the spectrum of the Gram matrix. [**?**] 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 correspond 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 $\boldsymbol{x} = (x_1, \dots, x_m)$ 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 [**?**]

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

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

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

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

How to improve stability ?

- by adding $n - m$ i.i.d. samples from μ, which corresponds to **volume sampling** [?]

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

  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** [?]

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

  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)**[1].

Theorem (N. and Michel 2023)

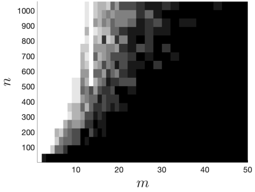
Assume that $(x_1, \dots, x_n)$ is drawn (by rejection) from $\gamma_m^{\otimes(n/m)}$ conditioned to the event $S_\delta = \{\lambda_{min}(\boldsymbol{G}) \geq 1 - \delta\}$. Then the weighted LS projection satisfies

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

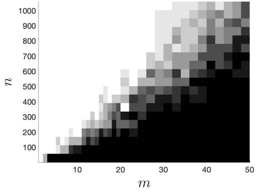
Similar theoretical guarantees as optimal i.i.d., but better concentration properties in practice.

[1] A. Nouy, B. Michel. Weighted least-squares approximation with determinantal point processes and generalized volume sampling. arXiv:2312.14057

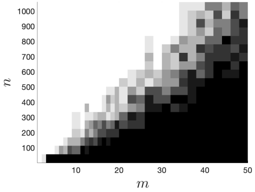
$\mathbb{P}(Sp(\boldsymbol{G}) \subset [1/2, 3/2])$ as a function of m and n



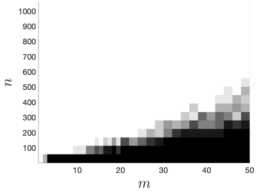
(a) i.i.d. μ (Classical)



(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 $[-1, 1]$.

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 Voigtlaender 2021] and tensor networks for i.i.d. samples [**?**].
- Quasi-optimality can be proven with i.i.d. sampling provided some condition $n \gtrsim K_w(M)$, which yields an optimal sampling strategy (only depending on M), but with unreasonable sampling complexity when M is a highly nonlinear manifold.
- More assumptions on functions are needed and algorithms and sampling should (in general) be adaptive.

A natural gradient descent

Consider a differentiable manifold M in the Hilbert space V and a natural gradient algorithm (in V) for solving

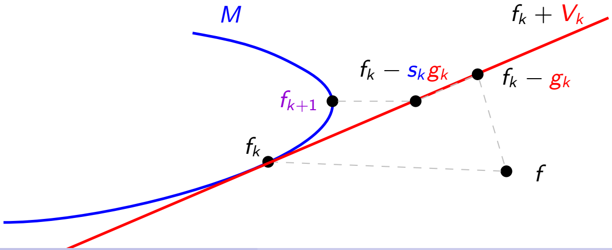
$$\inf_{v \in M} \mathcal{L}(v), \quad \mathcal{L}(v) := \frac{1}{2}\|f - v\|^2$$

which 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 $\nabla \mathcal{L}(f_k) = f_k - f$ onto V_k
- s_k a step size
- R_k a retraction map with values in M



Optimal sampling for natural gradient descent[2]

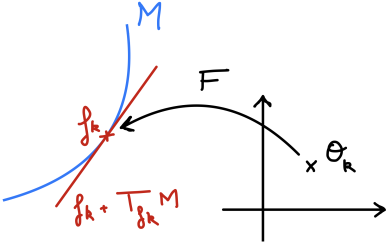
- g_k is defined as an empirical (quasi-)projection of the gradient onto V_k

$$g_k = \hat{P}_{V_k}(f_k - f)$$

  using evaluations of $f_k - f$ 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 $f_k = F(\theta_k)$,

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



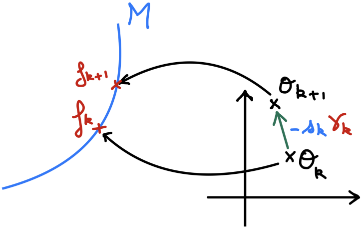
  or a subspace of $T_{f_k} M$.

[2] 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

$$R_k(f_k - s_k g_k) = F(\theta_k - s_k \gamma_k) \quad \text{for} \quad g_k(x) = \psi(x)^T \gamma_k.$$



- For $V = L^2_\mu$, taking

$$\gamma_k = (\psi, f_k - f)_n = \frac{1}{n} \sum_{i=1}^{n} \psi(x_i)(f_k(x_i) - f(x_i)) = \nabla_\theta(\mathcal{L}_n(F(\theta_k)))$$

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 $f_k - f$.
- Our algorithm can be seen as an preconditioned SGD using optimal sampling strategy.
- Convergence results are obtained for general risk functionals under classical smoothness and convexity assumptions on $\mathcal{L}$.

Convergence analysis

We make the following asumptions

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

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

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

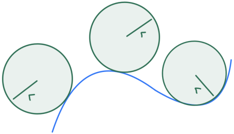
  Satisfied by (unbiased) quasi-projection or least-squares projections using i.i.d. samples from optimal distribution or (repeated) determinantal point processes. Requires a number of samples $n \lesssim d \log(d)$.

- 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.

Neural networks

We consider RePU shallow networks with width $s = 20$

$$M = \{F(\theta) = a^T \sigma(Ax + b) : \theta = (a, A, b) \in \mathbb{R}^s \times \mathbb{R}^{s \times d} \times \mathbb{R}^s\}, \quad \sigma(\cdot) = < \cdot >_+^2$$

for the approximation of $f(x) = \sin(2\pi x)$ on $[-1, 1]$.

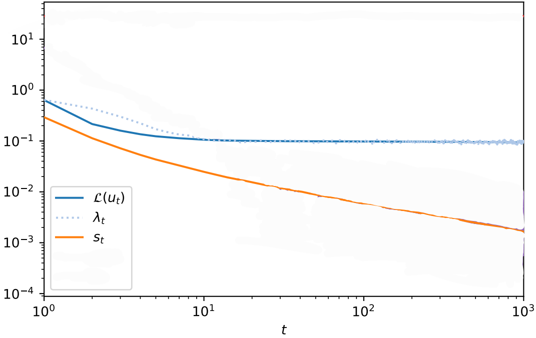


Figure: Loss $\mathcal{L}(u_k)$ for **SGD with classical sampling** and deterministically decreasing step sizes, plotted against the number of steps

Neural networks

We consider RePU shallow networks with width $s = 20$

$$M = \{F(\theta) = a^T \sigma(Ax + b) : \theta = (a, A, b) \in \mathbb{R}^s \times \mathbb{R}^{s\times d} \times \mathbb{R}^s\}, \quad \sigma(\cdot) = <\cdot>_+^2$$

for the approximation of $f(x) = \sin(2\pi x)$ on $[-1, 1]$.

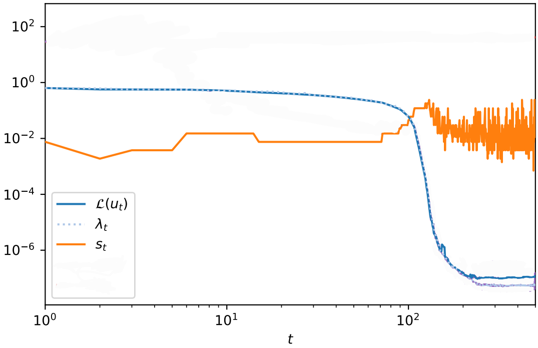


Figure: Loss $\mathcal{L}(u_k)$ for **NGD with optimal sampling**, least squares projection and adaptive step sizes, plotted against the number of steps

Outline

1. Optimal sampling for linear approximation
2. Optimal sampling for nonlinear approximation
3. More about linear approximation

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

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

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

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

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

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 [N. and Michel 2023].

Almost sure quasi-optimality in RKHS[3]

When V is a RKHS with kernel k, almost sure quasi-optimality in V-norm can be obtained by modifying the least-squares projection

$$\hat{f}_m = \arg\min_{v \in V_m} \|f - v\|_n^2, \quad \|f\|_n^2 = f(\boldsymbol{x})^T \boldsymbol{K}(\boldsymbol{x})^{-1} f(\boldsymbol{x})$$

$$\boldsymbol{x} = (x_1, \dots, x_n), \quad \boldsymbol{K}(\boldsymbol{x}) := (k(x_i, x_j))_{1 \le i,j \le n}$$

Letting $P_{V_{\boldsymbol{x}}}$ be the V-orthogonal projection onto $V_{\boldsymbol{x}} := span\{k(\cdot, x_i) : 1 \le i \le n\}$, it holds almost surely

$$\|f\|_n = \|P_{V_{\boldsymbol{x}}} f\|_V \le \|f\|_V$$

and the quasi-optimality

$$\|f - \hat{f}_m\|_V^2 \le (1 + \lambda_{min}(\boldsymbol{G}(\boldsymbol{x}))^{-1}) \inf_{v \in V_m} \|f - v\|_V^2$$

with the Gram matrix $\boldsymbol{G}(\boldsymbol{x}) = \varphi(\boldsymbol{x})^T \boldsymbol{K}(\boldsymbol{x})^{-1} \varphi(\boldsymbol{x})$, where $\varphi(x)$ is a V-orthonormal basis.

A sampling scheme should be chosen such that $\lambda_{min}(\boldsymbol{G}(\boldsymbol{x}))$ is controlled with high probability with a small sample size.

[3] P. Trunschke and A. Nouy. Almost-sure quasi-optimal approximation in reproducing kernel Hilbert spaces. arXiv:2407.06674.

Almost sure quasi-optimality in RKHS[4]

Continuous volume sampling [?] comes with theoretical guarantees

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

A better performance (without theoretical guarantees) is obtained with a subspace-informed volume sampling

$$det(\boldsymbol{G}(\boldsymbol{x}))d\nu^{\otimes n}(\boldsymbol{x})$$

Since $\lambda_{max}(\boldsymbol{G}(\boldsymbol{x})) \leq 1$, maximising $\det(\boldsymbol{G}(\boldsymbol{x}))$ allows a good control of $\lambda_{min}(\boldsymbol{G}(\boldsymbol{x}))$.

For $n = m$,

$$\det(\boldsymbol{G}(\boldsymbol{x})) = \frac{\det(\boldsymbol{\varphi}(\boldsymbol{x})^T \boldsymbol{\varphi}(\boldsymbol{x}))}{\det(k(\boldsymbol{x}, \boldsymbol{x}))}$$

which is a ratio of densities of determinantal point processes for V_m and H.

For $1 \leq i \leq m$, no explicit formula for conditional measures of x_i knowing $(x_1, \ldots, x_{i-1})$ but possible rejection algorithm.
For $i > m$, explicit formula for conditional distributions of of x_i knowing $(x_1, \ldots, x_{i-1})$.

[4] P. Trunschke and A. Nouy. Almost-sure quasi-optimal approximation in reproducing kernel Hilbert spaces. arXiv:2407.06674.

Conclusions

- Linear approximation using optimal i.i.d. or generalized volume sampling, and subsampling. Quasi-optimality with a low number of samples [1,2].
- Natural gradient method for nonlinear approximation using optimal sampling for linear approximation, with convergence guarantees [3].

  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). Possible sequential sampling strategy for a linear space defined by an arbitrary generating system [4].
- Still some computational challenges for general nonlinear classes (deep networks) and risk functionals.

---

[1] A. Nouy, B. Michel. Weighted least-squares approximation with determinantal point processes and generalized volume sampling. arXiv:2312.14057.
[2] P. Trunschke and A. Nouy. Almost-sure quasi-optimal approximation in reproducing kernel Hilbert spaces. arXiv:2407.06674.
[3] R. Gruhlke, A. Nouy and P. Trunschke. Optimal sampling for stochastic and natural gradient descent: arXiv:2402.03113.
[4] P. Trunschke and A. Nouy. Optimal sampling for least squares approximation with general dictionaries. arXiv:2407.07814.

References I

A. Cohen and G. Migliorati.
Optimal weighted least-squares methods.
SMAI Journal of Computational Mathematics, 3:181–203, 2017.

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 L_2.
arXiv e-prints, arXiv:2204.12621, Apr. 2022.

A. Chkifa and M. Dolbeault.
Randomized least-squares with minimal oversampling and interpolation in general spaces.
arXiv preprint arXiv:2306.07435, 2023.

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.

References II

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.

G. Migliorati.
Adaptive approximation by optimal weighted least-squares methods.
SIAM Journal on Numerical Analysis, 57(5):2217–2245, 2019.

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.

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.

References III

N. Nagel, M. Schäfer, and T. Ullrich.
A new upper bound for sampling numbers.
Foundations of Computational Mathematics, pages 1–24, 2021.

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.

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.

References IV

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.

M. Ali and A. Nouy.
Approximation theory of tree tensor networks: Tensorized univariate functions.
Constructive Approximation, 2023.

C. Haberstich, A. Nouy, and G. Perrin.
Active learning of tree tensor networks using optimal least-squares.
SIAM/ASA Journal on Uncertainty Quantification 11 (3), 848-876, 2023.

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.

A. Uschmajew and B. Vandereycken.
The geometry of algorithms using hierarchical tensors.
Linear Algebra and its Applications, 439(1):133–166, 2013.

B. Michel and A. Nouy.
Learning with tree tensor networks: Complexity estimates and model selection.
Bernoulli, 28(2):910 – 936, 2022.

References V

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.

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.

P. Trunschke.

Convergence bounds for nonlinear least squares for tensor recovery.

arXiv preprint arXiv:2208.10954, 2022.

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.