Optimal sampling for linear and nonlinear approximation

Anthony Nouy

Centrale Nantes, Nantes Université,
Laboratoire de Mathématiques Jean Leray

Joint works with
Robert Gruhlke, Cécile Haberstich, Bertrand Michel,
Guillaume Perrin, 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

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

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

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

- Sobolev or Besov smoothness: splines or wavelets
- Analytic smoothness: polynomials
- For a broader class of functions: tensor networks, neural networks
- Low-dimensional space or manifold $V_m = \{F(\theta) : \theta \in \mathbb{R}^m\}$ that approximate K, obtained by manifold approximation (or model order reduction) methods.

Approximation from limited information

In practice, an approximation $\hat{f}_m$ in V_m is constructed by an algorithm using only a limited number of information $\ell_1(f), \ldots, \ell_n(f)$, such as pointwise evaluations $f(x_1), \ldots, f(x_n)$ (standard information).

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

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

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

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

Approximation from limited information

In practice, an approximation $\hat{f}_m$ in V_m is constructed by an algorithm using only a limited number of information $\ell_1(f), \dots, \ell_n(f)$, such as pointwise evaluations $f(x_1), \dots, f(x_n)$ (standard information).

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

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

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

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

- 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 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

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

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 $\nu = w\mu$, 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

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

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

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

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

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

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

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

If the x_i are samples from $\nu = w\mu$,

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

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

$$K_w(V_m) = \sup_{x \in \mathcal{X}} w(x)^{-1} \|\boldsymbol{\varphi}(x)\|_2^2.$$

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

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

and an optimal sampling measure (leverage score sampling) is given by

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

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

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

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

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

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

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

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

Assume that $(x_1, \ldots, x_n)$ *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*

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

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

Subsampling

Subsampling methods start with a stable empirical Gram matrix obtained with $m \log(m)$ samples and select a (hopefully small) subsample preserving stability.

- In [Haberstich, N. and Perrin 2022][1], deterministic greedy subsampling algorithm:

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

  Often returns a number of samples close (or even equal) to m, without theoretical guaranty to downsample to $O(m)$.

- In [Dolbeault and Cohen 2022], subsampling algorithm based on successive random partitioning of the samples:

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

  with number of samples in $O(m)$, but not computationally feasible.

- In [Bartel, Schafer and T. Ullrich 2023], feasible subsampling algorithms ensuring $\lambda_{min}(\boldsymbol{G}) \geq 1 - \delta$ with $O(m)$ samples, but no guaranty of quasi-optimality in expectation.

[1] C. Haberstich, A. Nouy, and G. Perrin. Boosted optimal weighted least-squares. *Mathematics of Computation*, 91(335):1281–1315, 2022.

Introducing dependence

A 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 [Goreinov et al 2010, Fonarev et al 2016].

Introducing dependence

A 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 [Goreinov et al 2010, Fonarev et al 2016].

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

A 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 [Goreinov et al 2010, Fonarev et al 2016].

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 $\nu_m^{\otimes m}$.

It is a **projection determinantal point process (DPP)** for V_m [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 marginals are all equal to the optimal measure ν_m for i.i.d. sampling.

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 $\boldsymbol{\varphi}(x_i)$.

Projection DPP

From base-height formula of the determinant

$$\frac{1}{m!}\det(\boldsymbol{\varphi}(\boldsymbol{x})^T\boldsymbol{\varphi}(\boldsymbol{x})) = \underbrace{\frac{1}{m}\|\boldsymbol{\varphi}(x)\|_2^2}_{\sim x_1} \dots \underbrace{\frac{1}{m-k}\|\boldsymbol{\varphi}(x) - P_{W_k}\boldsymbol{\varphi}(x)\|_2^2}_{\sim x_{k+1}|x_1 \dots, x_k} \dots \underbrace{\|\boldsymbol{\varphi}(x) - P_{W_{m-1}}\boldsymbol{\varphi}(x)\|_2^2}_{\sim x_m|x_1 \dots x_{m-1}}$$

where P_{W_k} is the orthogonal projection onto the subspace

$$W_k = span\{\boldsymbol{\varphi}(x_1), \dots, \boldsymbol{\varphi}(x_k)\} \subset \mathbb{R}^m.$$

A sample $(x_1, \dots, x_m)$ from γ_m can be obtained by a sequential procedure

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

Projection DPP

From base-height formula of the determinant

$$\frac{1}{m!}\det(\boldsymbol{\varphi}(\boldsymbol{x})^T\boldsymbol{\varphi}(\boldsymbol{x})) = \underbrace{\frac{1}{m}\|\boldsymbol{\varphi}(x)\|_2^2}_{\sim x_1}\dots\underbrace{\frac{1}{m-k}\|\boldsymbol{\varphi}(x) - P_{W_k}\boldsymbol{\varphi}(x)\|_2^2}_{\sim x_{k+1}|x_1\dots,x_k}\dots\underbrace{\|\boldsymbol{\varphi}(x) - P_{W_{m-1}}\boldsymbol{\varphi}(x)\|_2^2}_{\sim x_m|x_1\dots x_{m-1}}$$

where P_{W_k} is the orthogonal projection onto the subspace

$$W_k = span\{\boldsymbol{\varphi}(x_1),\dots,\boldsymbol{\varphi}(x_k)\} \subset \mathbb{R}^m.$$

A sample $(x_1,\dots,x_m)$ from γ_m can be obtained by a sequential procedure

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

This is a randomized version of empirical interpolation

$$x_{k+1} = \arg\max_x \|\boldsymbol{\varphi}(x) - P_{W_k}\boldsymbol{\varphi}(x)\|_2^2$$

Projection DPP

From base-height formula of the determinant

$$\frac{1}{m!}\det(\varphi(\boldsymbol{x})^T\varphi(\boldsymbol{x})) = \underbrace{\frac{1}{m}\|\varphi(x)\|_2^2}_{\sim x_1} \dots \underbrace{\frac{1}{m-k}\|\varphi(x) - P_{W_k}\varphi(x)\|_2^2}_{\sim x_{k+1}|x_1\dots,x_k} \dots \underbrace{\|\varphi(x) - P_{W_{m-1}}\varphi(x)\|_2^2}_{\sim x_m|x_1\dots x_{m-1}}$$

where P_{W_k} is the orthogonal projection onto the subspace

$$W_k = span\{\varphi(x_1),\dots,\varphi(x_k)\} \subset \mathbb{R}^m.$$

A sample $(x_1,\dots,x_m)$ from γ_m can be obtained by a sequential procedure

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

This is a randomized version of empirical interpolation

$$\begin{aligned} x_{k+1} &= \arg\max_x \|\varphi(x) - P_{W_k}\varphi(x)\|_2^2 \\ &= \arg\max_x k(x,x) - k(x,\underline{x})k(\underline{x},\underline{x})^{-1}k(\underline{x},x), \quad \underline{x} = (x_1,\dots,x_k) \end{aligned}$$

or adaptive gaussian process interpolation with projection kernel $k(x,y) = \varphi(x)^T\varphi(y)$.

Improving stability

Stability can be ensured with high probability

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

  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.

Improving stability

Stability can be ensured with high probability

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

  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.

- by using multiple samples from γ_m **(repeated DPP)**.

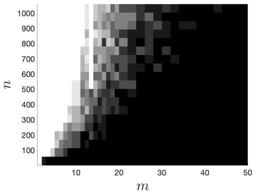
Theorem (N. and Michel 2023)

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 least-squares projection satisfies

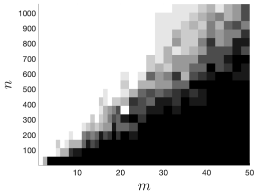
$$\mathbb{E}(\|f - \hat{f}_m\|^2) \leq (1 + \frac{m}{n}\mathbb{P}(S_\delta)^{-1}(1-\delta)^{-2}) \inf_{g \in V_m} \|f - g\|^2.$$

Similar theoretical results as for i.i.d., but better concentration properties in practice.

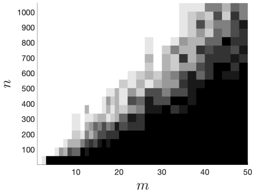
$\mathbb{P}(Sp(\boldsymbol{G}) \subset [1/2, 3/2])$ as a function of m and n



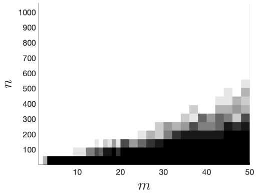
(a) i.i.d. μ (Classical)



(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 $[-1, 1]$.

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.

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 [Eigel, Schneider and Trunschke, 2022].

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 [Eigel, Schneider and Trunschke, 2022].

- Quasi-optimality can be proven with i.i.d. sampling provided

  $$n \gtrsim K_w(M) = \sup_{x \in \mathcal{X}} w(x)^{-1} \kappa_M(x) \quad (\kappa_M^{-1} : \text{Generalized Christoffel function})$$

  that yields an optimal i.i.d. sampling strategy [Trunschke 2022, Cardenas et al 2024]

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 [Eigel, Schneider and Trunschke, 2022].

- Quasi-optimality can be proven with i.i.d. sampling provided

  $$n \gtrsim K_w(M) = \sup_{x \in \mathcal{X}} w(x)^{-1} \kappa_M(x) \quad (\kappa_M^{-1} : \text{Generalized Christoffel function})$$

  that yields an optimal i.i.d. sampling strategy [Trunschke 2022, Cardenas et al 2024]
- However, in general, no real benefit compared to classical sampling. E.g. for sets M of low-rank tensors in a tensor space $U^{\otimes d}$, $K_w(M) = K_w(U^{\otimes})$, that yields the condition

  $$n \gtrsim \dim(U)^d \quad \text{(curse of dimensionality)}$$

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 [Eigel, Schneider and Trunschke, 2022].
- Quasi-optimality can be proven with i.i.d. sampling provided

  $$n \gtrsim K_w(M) = \sup_{x \in \mathcal{X}} w(x)^{-1} \kappa_M(x) \quad (\kappa_M^{-1} : \text{Generalized Christoffel function})$$

  that yields an optimal i.i.d. sampling strategy [Trunschke 2022, Cardenas et al 2024]
- However, in general, no real benefit compared to classical sampling. E.g. for sets M of low-rank tensors in a tensor space $U^{\otimes d}$, $K_w(M) = K_w(U^{\otimes})$, that yields the condition

  $$n \gtrsim \dim(U)^d \quad \text{(curse of dimensionality)}$$

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

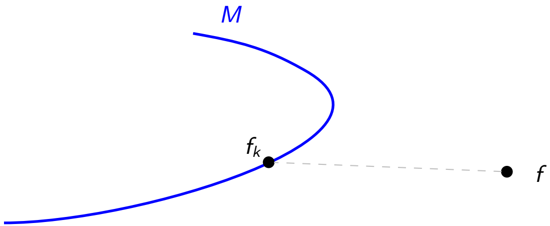
Active learning for natural gradient descent

Consider a differentiable manifold M and a natural gradient algorithm (in function space) for solving

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

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

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



Active learning for natural gradient descent

Consider a differentiable manifold M and a natural gradient algorithm (in function space) for solving

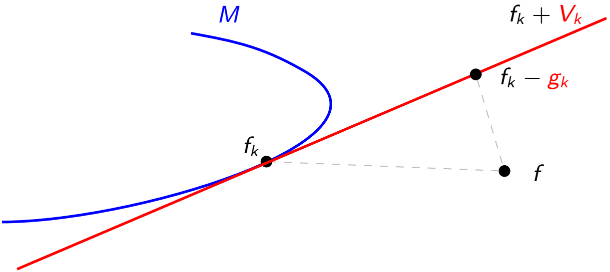
$$\inf_{v\in M} \mathcal{L}(v), \quad \mathcal{L}(v) := \|f - v\|^2$$

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

- V_k is a local approximation of $M - f_k$
- g_k a projection of the gradient $\nabla\mathcal{L}(f_k) = f_k - f$ onto V_k



Active learning for natural gradient descent

Consider a differentiable manifold M and a natural gradient algorithm (in function space) for solving

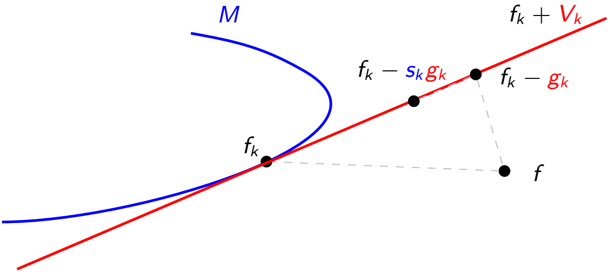
$$\inf_{v \in M} \mathcal{L}(v), \quad \mathcal{L}(v) := \|f - v\|^2$$

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

- V_k is a local approximation of $M - f_k$
- g_k a projection of the gradient $\nabla \mathcal{L}(f_k) = f_k - f$ onto V_k
- s_k a step size



Active learning for natural gradient descent

Consider a differentiable manifold M and a natural gradient algorithm (in function space) for solving

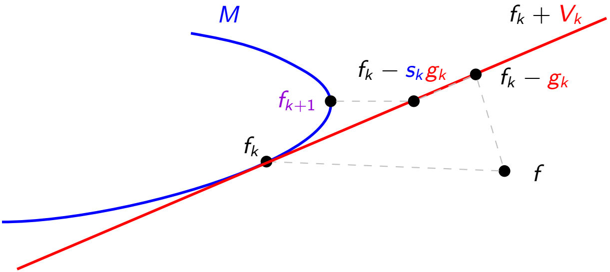
$$\inf_{v \in M} \mathcal{L}(v), \quad \mathcal{L}(v) := \|f - v\|^2$$

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

- V_k is a local approximation of $M - f_k$
- 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



Active learning for natural gradient descent

- 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.

Active learning for natural gradient descent

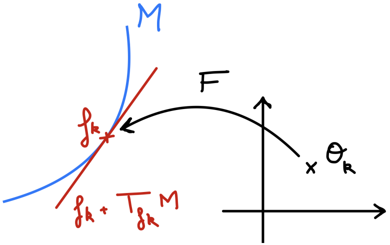
- g_k is defined as an empirical (quasi-)projection of the gradient onto V_k

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

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 $f_k = F(\theta_k)$,

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

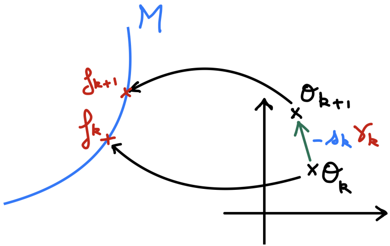


or a subspace of $T_{f_k}M$.

Active learning for natural gradient descent

- A natural (but not easy to control) retraction is

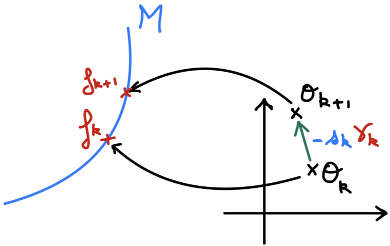
$$R_k(f_k - s_k g_k) = F(\theta_k - s_k \gamma_k) \quad \text{for} \quad g_k(x) = \psi(x)^T \gamma_k.$$



Active learning for natural gradient descent

- A natural (but not easy to control) retraction is

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



- Taking

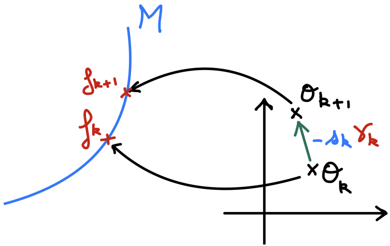
$$\gamma_k = (\psi, f_k - f)_n = \frac{1}{n}\sum_{i=1}^{n} \psi(x_i)(f_k(x_i) - f(x_i)) = \nabla_\theta(\mathcal{L}_n(F(\theta_k)))$$

corresponds to classical batch stochastic gradient descent (SGD), where g_k is a quasi-projection on V_k that can be very far from the orthogonal projection of $f_k - f$.

Active learning for natural gradient descent

- A natural (but not easy to control) retraction is

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



- 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 that 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 analysis

We make the following asumptions

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

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

  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 $n \lesssim d \log(d)$.

Convergence analysis

We make the following asumptions

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

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

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

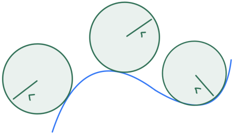
  Satisfied by (unbiased) quasi-projection or least-squares projections using i.i.d. samples from optimal distribution or (repeated) determinantal point processes. Requires a number of samples $n \lesssim d \log(d)$.

- The retraction map R_k at f_k satisfies

$$\|R_k(f_k + g)\|^2 \leq \|f_k + g - f\|^2 + \frac{C_R}{2}\|g\|^2 + \beta_k \qquad \text{(CR)}$$

  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

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

where

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

Convergence analysis

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

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

where

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

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

$$\mathbb{E}(\|f_{k+1} - f\|^2 | \mathcal{F}_k) \leq \mathbb{E}(\|f_k - f\|^2 | \mathcal{F}_k)$$

  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.

Convergence analysis

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

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

where

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

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

  $$\mathbb{E}(\|f_{k+1} - f\|^2 | \mathcal{F}_k) \leq \mathbb{E}(\|f_k - f\|^2 | \mathcal{F}_k)$$

  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 error $f - f^k$. Feasible with sufficiently small c_b (large n).

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

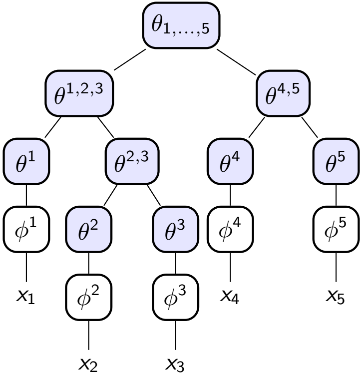
Tree tensor networks

Tree tensor networks form a prominent class of approximation tools for the approximation of multivariate functions $f(x_1, \dots, x_d)$. 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:

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



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.

[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.

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.

[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.

Tree tensor networks

M is a differentiable manifold[2] with tangent space

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

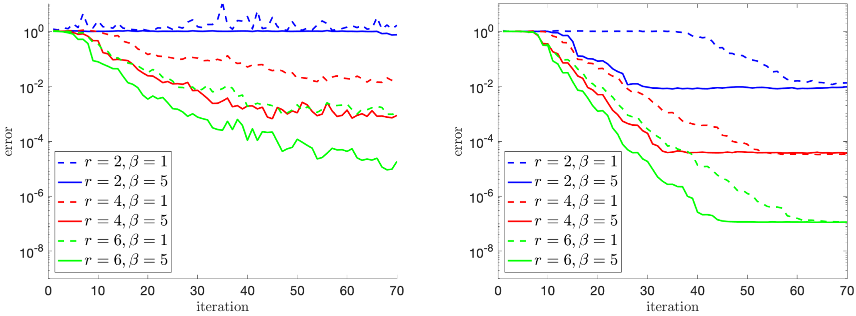
Controlled retraction using higher order singular value decomposition.

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.

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 $n = \beta 4d \log(4d)$, $d = \dim(V_k)$.

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 $[-1, 1]$.

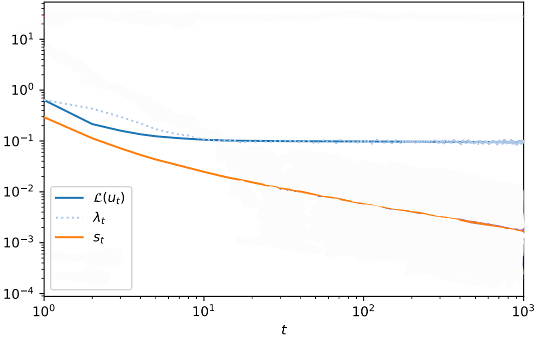


Figure: Loss $\mathcal{L}(u_k)$ for **SGD with classical sampling** and deterministically decreasing step sizes, plotted against the number of steps

Neural networks

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

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

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

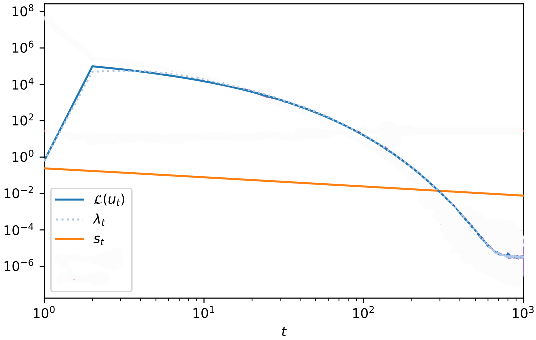


Figure: Loss $\mathcal{L}(u_k)$ for **NGD with optimal sampling**, least squares projection and deterministically decreasing step sizes, plotted against the number of steps

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 $[-1, 1]$.

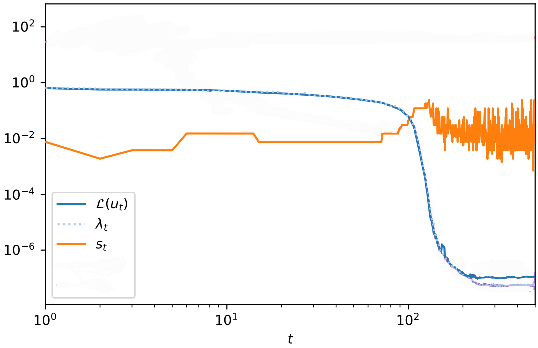


Figure: Loss $\mathcal{L}(u_k)$ for **NGD with optimal sampling**, least squares projection and adaptive step sizes, plotted against the number of steps

Outline

Sampling from general generating systems

Assume we have access to a (non orthonormal) generating system $\boldsymbol{\psi} = (\psi_1, \dots, \psi_d)$ of a linear V_m, e.g. $\boldsymbol{\psi} = \nabla_\theta F(\theta)$ for $M = \{F(\theta) : \theta \in \mathbb{R}^d\}$.

Optimal sampling density for V_m is given by

$$w_\star(x) = \frac{1}{m}\|\boldsymbol{\varphi}(x)\|_2^2 = \frac{1}{m}\boldsymbol{\psi}(x)^T \boldsymbol{G}_\star^+ \boldsymbol{\psi}(x),$$

where $\boldsymbol{G}_\star$ is the Gram matrix of $\boldsymbol{\psi}$.

[3] A. Nouy and P. Trunschke. Optimal sampling for least squares approximation with general dictionaries, coming soon.

Sampling from general generating systems

Assume we have access to a (non orthonormal) generating system $\boldsymbol{\psi} = (\psi_1, \dots, \psi_d)$ of a linear V_m, e.g. $\boldsymbol{\psi} = \nabla_\theta F(\theta)$ for $M = \{F(\theta) : \theta \in \mathbb{R}^d\}$.

Optimal sampling density for V_m is given by

$$w_\star(x) = \frac{1}{m}\|\boldsymbol{\varphi}(x)\|_2^2 = \frac{1}{m}\boldsymbol{\psi}(x)^T \boldsymbol{G}_\star^+ \boldsymbol{\psi}(x),$$

where $\boldsymbol{G}_\star$ is the Gram matrix of $\boldsymbol{\psi}$.

An approximately orthogonal basis can be obtained from an estimate of the Gram matrix

$$\boldsymbol{G} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\psi}(x_i)\boldsymbol{\psi}(x_i)^T.$$

If n is sufficiently large to ensure

$$(1-\epsilon)\boldsymbol{G}_\star \le \boldsymbol{G} \le (1+\epsilon)\boldsymbol{G}_\star \quad \Longrightarrow \quad (1+\epsilon)^{-1} w_\star \le w \le (1-\epsilon)^{-1} w_\star$$

But this requires $n \gtrsim K_{1,m}$, which may grow exponentially with m or even be unbounded.

[3] A. Nouy and P. Trunschke. Optimal sampling for least squares approximation with general dictionaries, coming soon.

Sampling from general generating systems

Assume we have access to a (non orthonormal) generating system $\boldsymbol{\psi} = (\psi_1, \dots, \psi_d)$ of a linear V_m, e.g. $\boldsymbol{\psi} = \nabla_\theta F(\theta)$ for $M = \{F(\theta) : \theta \in \mathbb{R}^d\}$.

Optimal sampling density for V_m is given by

$$w_\star(x) = \frac{1}{m}\|\boldsymbol{\varphi}(x)\|_2^2 = \frac{1}{m}\boldsymbol{\psi}(x)^T \boldsymbol{G}_\star^+ \boldsymbol{\psi}(x),$$

where $\boldsymbol{G}_\star$ is the Gram matrix of $\boldsymbol{\psi}$.

An approximately orthogonal basis can be obtained from an estimate of the Gram matrix

$$\boldsymbol{G} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\psi}(x_i)\boldsymbol{\psi}(x_i)^T.$$

If n is sufficiently large to ensure

$$(1-\epsilon)\boldsymbol{G}_\star \le \boldsymbol{G} \le (1+\epsilon)\boldsymbol{G}_\star \quad \Longrightarrow \quad (1+\epsilon)^{-1} w_\star \le w \le (1-\epsilon)^{-1} w_\star$$

But this requires $n \gtrsim K_{1,m}$, which may grow exponentially with m or even be unbounded.

A bootstrap strategy can be used, with convergence guarantees[3]

$$\boldsymbol{G}_{k+1} = \frac{k}{k+1}\boldsymbol{G}_k + \frac{1}{k}\boldsymbol{H}_k, \quad \boldsymbol{H}_k = \frac{1}{n}\sum_{i=1}^n w_k(x_i)^{-1}\boldsymbol{\psi}(x_i)\boldsymbol{\psi}(x_i)^T, \quad x_i \sim w_k \mu$$

[3] A. Nouy and P. Trunschke. Optimal sampling for least squares approximation with general dictionaries, coming soon.

More general metrics... towards physics informed optimal sampling

Consider a Hilbert space V of functions defined on $\mathcal{X}$ equipped with the norm

$$\|f\|^2 = \int_{\mathcal{X}} |L_x f|^2 d\mu(x), \quad L_x : \mathcal{H} \to \mathbb{R}^\ell \; (\textit{linear})$$

e.g. $V = L^2_\mu(\mathcal{X})$ for $L_x f = f(x)$ or $V = H^1_\mu(\mathcal{X})$ with $L_x f = \begin{pmatrix} f(x) \\ \nabla f(x) \end{pmatrix}$.

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

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

An optimal sampling measure for i.i.d. sampling is given by the density

$$w_m(x) = \alpha^{-1} \|L_x \boldsymbol{\varphi}\|_2^2, \quad L_x \boldsymbol{\varphi} \in \mathbb{R}^{m \times \ell},$$

with $\alpha \leq m$. With conditioned sampling and $\mathcal{O}(m \log(m))$ samples, we prove quasi-optimality result in expectation in the V norm[4].

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

Control in probability

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

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

This requires an 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 $g \in H$,

$$\|g\| \leq C_H \|g\|_H \quad \text{(continuous embedding } H \hookrightarrow L^2_\mu\text{)}$$

and

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

it holds almost surely

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

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

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

For example

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

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

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

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

For example

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

Then $\|g\|_n \leq 2\|g\|_H$ holds by choosing for the weight function a mixture

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

For i.i.d. sampling from $w\mu$, the empirical Gram matrix $\boldsymbol{G}$ remains an unbiased estimator of $\boldsymbol{I}$ and

$$K_{w,m} = \sup_{x\in\mathcal{X}} w(x)\|\boldsymbol{\varphi}(x)\|_2^2 \leq 2K_{w_m,m} = 2m$$

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.

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

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

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

For example

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

Then $\|g\|_n \leq 2\|g\|_H$ holds by choosing for the weight function a mixture

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

For i.i.d. sampling from $w\mu$, the empirical Gram matrix $\boldsymbol{G}$ remains an unbiased estimator of $\boldsymbol{I}$ and

$$K_{w,m} = \sup_{x\in\mathcal{X}} w(x)\|\varphi(x)\|_2^2 \leq 2K_{w_m,m} = 2m$$

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.[5]

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

Almost sure quasi-optimality in RKHS[6]

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

$$\hat{f}_m = \arg\min_{v \in V_m} \|f - v\|_n^2, \quad \|f\|_n^2 = f(\boldsymbol{x})^T k(\boldsymbol{x}, \boldsymbol{x})^{-1} f(\boldsymbol{x}), \quad \boldsymbol{x} = (x_1, \dots, x_n)$$

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

$$\|f\|_n = \|P_{H_{\boldsymbol{x}}} f\|_H \leq \|f\|_H$$

and the quasi-optimality

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

with the Gram matrix $\boldsymbol{G}(\boldsymbol{x}) = \varphi(\boldsymbol{x})^T k(\boldsymbol{x}, \boldsymbol{x})^{-1} \varphi(\boldsymbol{x})$.

$\lambda_{max}(\boldsymbol{G}(\boldsymbol{x})) \leq 1$ and sampling from $\det(\boldsymbol{G}(\boldsymbol{x}))$ allows to control $\lambda_{min}(\boldsymbol{G}(\boldsymbol{x}))$. For $n = m$,

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

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

[6] A. Nouy and P. Trunschke. Almost-sure quasi-optimal least squares approximation. Coming soon.

Conclusions

- Linear approximation using optimal i.i.d. or generalized volume sampling. Quasi-optimality with a low number of samples [1,2,3].

---

[1] C. Haberstich, A. Nouy, and G. Perrin. Boosted optimal weighted least-squares. Mathematics of Computation, 91(335):1281–1315, 2022.
[2] A. Nouy, B. Michel. Weighted least-squares approximation with determinantal point processes and generalized volume sampling. arXiv:2312.14057.
[3] A. Nouy, P. Trunschke. Almost-sure quasi-optimal least squares approximation. Coming soon.
[4] R. Gruhlke, A. Nouy and P. Trunschke. Optimal sampling for stochastic and natural gradient descent: arXiv:2402.03113.
[5] A. Nouy, P. Trunschke. Optimal sampling for least squares approximation with general dictionaries. Coming soon.

Conclusions

- Linear approximation using optimal i.i.d. or generalized volume sampling. Quasi-optimality with a low number of samples [1,2,3].
- Natural gradient method for nonlinear approximation in an active learning setting using optimal sampling for linear approximation. Convergence guarantees [4].

  Applies to a large class of risk functionals and metrics... towards physics informed optimal sampling and other machine learning tasks .

---

[1] C. Haberstich, A. Nouy, and G. Perrin. Boosted optimal weighted least-squares. Mathematics of Computation, 91(335):1281–1315, 2022.

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

[3] A. Nouy, P. Trunschke. Almost-sure quasi-optimal least squares approximation. Coming soon.

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

[5] A. Nouy, P. Trunschke. Optimal sampling for least squares approximation with general dictionaries. Coming soon.

Conclusions

- Linear approximation using optimal i.i.d. or generalized volume sampling. Quasi-optimality with a low number of samples [1,2,3].
- Natural gradient method for nonlinear approximation in an active learning setting using optimal sampling for linear approximation. Convergence guarantees [4].

  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 tree tensor networks and shallow networks in L^2 setting. Possible sequential sampling strategy for a linear space defined by an arbitrary generating system [5].

---

[1] C. Haberstich, A. Nouy, and G. Perrin. Boosted optimal weighted least-squares. Mathematics of Computation, 91(335):1281–1315, 2022.

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

[3] A. Nouy, P. Trunschke. Almost-sure quasi-optimal least squares approximation. Coming soon.

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

[5] A. Nouy, P. Trunschke. Optimal sampling for least squares approximation with general dictionaries. Coming soon.

Conclusions

- Linear approximation using optimal i.i.d. or generalized volume sampling. Quasi-optimality with a low number of samples [1,2,3].
- Natural gradient method for nonlinear approximation in an active learning setting using optimal sampling for linear approximation. Convergence guarantees [4].

  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 tree tensor networks and shallow networks in L^2 setting. Possible sequential sampling strategy for a linear space defined by an arbitrary generating system [5].
- Still some computational challenges for general nonlinear classes (deep networks) and risk functionals.

---

[1] C. Haberstich, A. Nouy, and G. Perrin. Boosted optimal weighted least-squares. Mathematics of Computation, 91(335):1281–1315, 2022.

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

[3] A. Nouy, P. Trunschke. Almost-sure quasi-optimal least squares approximation. Coming soon.

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

[5] A. Nouy, P. Trunschke. Optimal sampling for least squares approximation with general dictionaries. Coming soon.

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.

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.

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.

References II

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.

N. Nagel, M. Schäfer, and T. Ullrich.

A new upper bound for sampling numbers.

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

References III

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.

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.

References IV

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.

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.

References V

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.