Kernelization of Natural Gradient Methods for Physics Informed Neural Network (PINNs)

MIA Paris-Saclay Seminar, AgroParisTech-INRAE

Nilo Schwencke

Formerly: PhD @ TAU/A&O Team (INRIA; LISN, Paris-Saclay University; CNRS).
Soon: Postdoc @ OCKHAM Team (INRIA; LIP, ENS Lyon; CNRS)

January 22, 2026

Scientific Machine-Learning (SciML) and Physics Informed Neural Networks (PINNs)

- **Data-driven modeling**: learn governing equations directly from observed trajectories (e.g., SINDy (Brunton et al., 2016), Neural ODEs (Chen et al., 2018)).
- **Operator learning**: approximate full solution operators for families of PDEs (e.g., DeepONet (Lu et al., 2021), FNO (Li et al., 2020)).
- **Physics-informed ML** : embed physical laws as constraints in the learning process (e.g., PINNs and extensions).

PIML aims to minimize:

$$\mathcal{L}(u) := \int_{\Omega} \|D[u] - f\|^2_{\mathbb{R}^n} + \int_{\partial\Omega} \|B[u] - g\|^2_{\mathbb{R}^m} .$$

PINNs key idea

- model u with a neural network (Lagaris et al., 1998)
- use autodiff to compute D and B (Raissi et al., 2019)

Sampling yields the loss:

$$\widehat{\ell}_{D,B}(\theta) := \frac{1}{2S_D} \sum_{i=1}^{S_D} \left(D[u_\theta](x_i^D) - f(x_i^D) \right)^2 + \frac{1}{2S_B} \sum_{i=1}^{S_B} \left(B[u_\theta](x_i^B) - g(x_i^B) \right)^2 .$$

Problem: This leads to low accuracy when using usual optimizers.

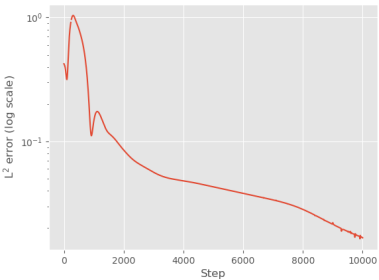


Figure: L^2-error of a PINN optimized with Adam on the 2D Laplace equation.

PINNs main drawback : optimization error

- Highly nonconvex & ill-conditioned optimization
- Gradient imbalance between loss terms
- Spectral bias of neural nets : Low frequencies learned first; high-frequency or localized features missed.

Intuition from Fourier

$$S_N : (\alpha_k) \in \mathbb{C}^{[\![-N,N]\!]} \mapsto \sum_{k=-N}^{N} \alpha_k e^{2i\pi kx}.$$

S_N singular values are all 1. **BUT:**

$$\Delta[S_N] \text{ spectrum is } \{4\pi^2 k^2 : 1 \leqslant k \leqslant N\}$$

Δ strongly impact the spectral conditionning.

Natural Gradient

Classical quadratic regression problem, with batch (x_i):

$$\widehat{\ell}(\theta) := \frac{1}{2S} \sum_{i=1}^{S} \left(u_\theta(x_i) - f(x_i)\right)^2 .$$

In the population limit:

$$\widehat{\ell}(\theta) \overset{S\to\infty}{\to} \mathcal{L}(u_\theta) ; \quad \mathcal{L}(u) := \frac{1}{2} \|u - f\|^2_{\mathrm{L}^2(\Omega)}$$

This yields the Fréchet derivative:

$$\mathrm{d}\mathcal{L}|_u(h) = \langle \underbrace{u - f}_{\nabla \mathcal{L}_{|u}}, h \rangle_{\mathrm{L}^2(\Omega)},$$

and thus the gradient flow:

$$\begin{cases} u_0 \in \mathrm{L}^2(\Omega) \\ \dot{u}_t = -\nabla \mathcal{L}_{|u_t} = f - u_t \end{cases} .$$

Solution: $u_t = f - e^{-t}(f - u_0)$.

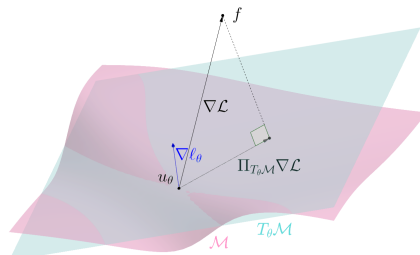
Natural gradient in functional space

The functional space is constrained to:

- $\mathcal{M} := \mathrm{Im}\, u = \{u_\theta \,:\, \theta \in \mathbb{R}^P)$
- $T_\theta \mathcal{M} := \mathrm{Im}\, \mathrm{d}u_\theta = \mathrm{Span}(\partial_p u_\theta)$

The Natural Gradient is then (Amari and Douglas, 1998):

$$\theta_{t+1} \leftarrow \theta_t - \eta\, \mathrm{d}u^\dagger_{\theta_t} \left(\Pi^\perp_{T_{\theta_t}\mathcal{M}} \nabla \mathcal{L}_{|u_{\theta_t}} \right),$$



Definition-Proposition

*The **Natural Neural Tangent Kernel (NNTK)** is the kernel of the projection $\Pi_{T_\theta \mathcal{M}} : L^2(\Omega) \to L^2(\Omega)$ onto $T_\theta \mathcal{M}$. It is given by the formula:*

$$NNTK_\theta(x, y) := \sum_{1 \leqslant p,q \leqslant P} (\partial_p u_\theta(x)) \quad G^\dagger_{\theta\, pq} \, (\partial_q u_\theta(y))^t ; \quad G_{\theta\, p,q} \quad := \langle \partial_p u_\theta \, , \, \partial_q u_\theta \rangle_{L^2(\Omega)} .$$

Corollary

The Natural Gradient update rewrites: $\theta_{t+1} \leftarrow \theta_t - \eta \, G^\dagger_{\theta_t} \nabla \ell(\theta_t); \quad \ell(\theta) := \mathcal{L}(u_\theta).$

Shortcomings

- Computation of the Gram matrix G_{θ_t} is quadratic in the number of parameters.
- Inversion of G_{θ_t} is cubic

We introduce the empirical Natural Gradient that scales linearly with the number of parameters.

Reproducing Kernel Hilbert Spaces (RKHS) *détour*

Definition-Proposition

An Hilbert space $\mathcal{H}$ of functions $\Omega \to \mathbb{R}$ is a RKHS if and only if the following equivalent conditions are met:

1. *There exist a function $k : \Omega \times \Omega \to \mathbb{R}$ such that:*
   - $\mathcal{H} = \overline{\text{Span}\,(k(x,\cdot) : x \in \Omega)}$
   - $\langle k(x,\cdot)\,,\, k(y,\cdot\rangle_{\mathcal{H}} = k(x,y)$
2. *for all $x \in \Omega$, the evaluation form $e_x : f \in \mathcal{H} \mapsto f(x)$ is continuous.*

Proposition

Any finite dimensional space $\mathcal{H}$ is a RKHS

Proposition

If $\mathcal{H} := \overline{\text{Span}(u_i : i \in \mathbb{N})}$, is an RKHS, then its kernel is given by: for all $x, y \in \Omega$

$$k(x,y) = \sum_{i,j \in \mathbb{N}} u_i(x) G^{\dagger}_{i,j} u_j(y)$$

where $G_{ij} := \langle u_i\,,\, u_j\rangle_{\mathcal{H}}$.

Proposition

If $\mathcal{H} \supset \mathcal{H}_0$ is an RKHS with kernel k, then the orthogonal projection $\Pi^{\perp}_{\mathcal{H}_0}$ onto $\mathcal{H}_0$ is given by:

$$\Pi^{\perp}_{\mathcal{H}_0}(f)(x) = \langle k(x,\cdot)\,,\, f\rangle_{\mathcal{H}}$$

Remark

Applying to $\mathcal{H}_0 = T_\theta \mathcal{M}$, we get natural gradient.

Intuition

$$\widehat{\ell}(u) := \frac{1}{2S} \sum_{i=1}^{S} \left(u(x_i) - f(x_i)\right)^2.$$

$$\frac{\mathrm{d}u_{\theta(t)}}{\mathrm{d}t}(x) = -\frac{1}{S} \sum_{i=1}^{S} \left(u_\theta(x_i) - f(x_i)\right) \delta_{x_i}$$

Gradient descent (Jacot et al., 2018)

$$\frac{\mathrm{d}u_{\theta(t)}}{\mathrm{d}t}(x) = -\frac{1}{S} \sum_{i=1}^{S} \mathrm{NTK}_\theta(\mathrm{x}, \mathrm{x_i}) \left(\mathrm{u}_\theta(\mathrm{x_i}) - \mathrm{f}(\mathrm{x_i})\right)$$

$$\text{with: } \mathrm{NTK}_\theta(\mathrm{x}, \mathrm{y}) := \sum_{\mathrm{p}=1}^{\mathrm{P}} \partial_{\mathrm{p}} \mathrm{u}_\theta(\mathrm{x}) \partial_{\mathrm{p}} \mathrm{u}_\theta(\mathrm{y})^\top.$$

Natural Gradient (Rudner et al., 2019)

$$\frac{\mathrm{d}u_{\theta(t)}}{\mathrm{d}t}(x) = -\frac{1}{S} \sum_{i=1}^{S} \mathrm{NNTK}_\theta(\mathrm{x}, \mathrm{x_i}) \left(\mathrm{u}_\theta(\mathrm{x_i}) - \mathrm{f}(\mathrm{x_i})\right)$$

$$\text{with: } \mathrm{NNTK}_\theta(\mathrm{x}, \mathrm{y}) := \sum_{1 \leqslant \mathrm{p}, \mathrm{q} \leqslant \mathrm{P}} \partial_{\mathrm{p}} \mathrm{u}_\theta(\mathrm{x}) \mathrm{G}^\dagger_{\theta_{\mathrm{p,q}}} \partial_{\mathrm{q}} \mathrm{u}_\theta(\mathrm{y})^\top,$$

$$G_{\theta_{p,q}} = \left\langle \partial_p u_\theta \, , \, \partial_q u_\theta \right\rangle.$$

Neural Tangent Kernel (NTK)

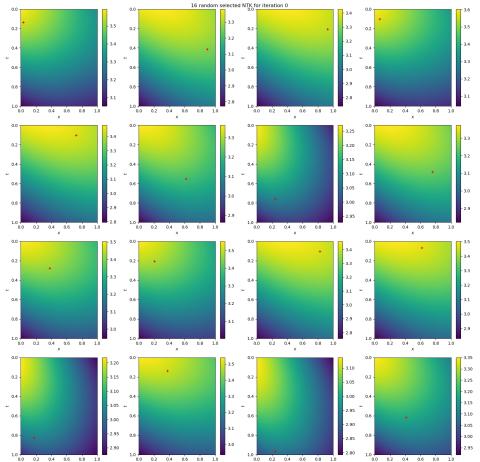


Figure: NTK for the Heat equation.

Reading: the plot represents the function $NTK_{\theta_0}(\cdot, x_i)$; the red cross marks the point x_i.

empirical Natural Gradient

Recall the the functional dynamics of (N)GD on the empirical loss $\widehat{\ell}$:

$$\frac{\mathrm{d}u_{\theta_t}}{\mathrm{d}t}(x) = -\sum_{i=1}^{S} (N)NTK_{\theta_t}(x, x_i)(u_{\theta_t}(x_i) - y_i),$$

Key Observation

The empirical dynamics takes place in:

$$\widehat{T}_\theta\mathcal{M} := \mathrm{Span}\left((N)NTK_\theta(x_i, \cdot) \,:\, (x_i)_{1\leqslant i\leqslant N}\right).$$

We can define the empirical Natural Gradient:

$$\theta_{t+1} = \theta_t - \eta\, \mathrm{d}u^\dagger_{\theta_t}\left(\Pi^\perp_{\widehat{T}_{\theta_t}\mathcal{M}}\nabla\mathcal{L}_{|u_{\theta_t}}\right).$$

Theorem (ANaGRAM)

Under mild assumptions:

$$du^\dagger_{\theta_t}\left(\Pi^\perp_{\widehat{T}_{\theta_t}\mathcal{M}}\nabla\mathcal{L}_{|u_{\theta_t}}\right) \simeq \widehat{\phi}^\dagger_{\theta_t}\widehat{\nabla\mathcal{L}}_{\theta_t},$$

with: for all $1 \leqslant p \leqslant P, 1 \leqslant i \leqslant S$

- $\widehat{\phi}_{\theta_t\, i,p} := \partial_p u_{\theta_t}(x_i)$
- $\widehat{\nabla\mathcal{L}}_{\theta_t\, i} := \nabla\mathcal{L}_{|u_{\theta_t}}(x_i)$

Key fact

$\hat{\phi}^\dagger_{\theta_t}$ **can be computed with a SVD, with complexity** $\mathrm{O}\big(\min(\mathrm{PS}^2, \mathrm{P}^2\mathrm{S})\big)$.

Corollary

There exists P points $(\hat{x}_i)$ *such that:*

$$\Pi^\perp_{\widehat{T}_\theta\mathcal{M}}\nabla\mathcal{L}_{|u_\theta} = \Pi^\perp_{T_\theta\mathcal{M}}\nabla\mathcal{L}_{|u_\theta}.$$

Recall the the functional dynamics of (N)GD on the empirical loss $\widehat{\ell}$:

$$\frac{\mathrm{d}u_{\theta_t}}{\mathrm{d}t}(x) = -\sum_{i=1}^{S} (N)NTK_{\theta_t}(x, x_i)(u_{\theta_t}(x_i) - y_i),$$

Key Observation

The empirical dynamics takes place in:

$$\widehat{T}_\theta\mathcal{M} := \mathrm{Span}\left((N)NTK_\theta(x_i, \cdot) \,:\, (x_i)_{1\leqslant i\leqslant N}\right).$$

We can define the empirical Natural Gradient:

$$\theta_{t+1} = \theta_t - \eta\, \mathrm{d}u^\dagger_{\theta_t}\left(\Pi^\perp_{\widehat{T}_{\theta_t}\mathcal{M}}\nabla\mathcal{L}_{|u_{\theta_t}}\right).$$

Byproduct

Yields an optimal criterion for (x_i) choice:

$$(x_i)^\star = \underset{(x_i)\in\Omega^S}{\mathrm{argmin}} \left\| \Pi^\perp_{\widehat{T}^{(x_i)}_{\theta,K}\mathcal{M}}\nabla\mathcal{L}_{|u_{\theta_t}} - \nabla\mathcal{L}_{|u_{\theta_t}} \right\|_{\mathrm{L}^2(\Omega)}$$

Theorem (ANaGRAM)

Under mild assumptions:

$$du^\dagger_{\theta_t}\left(\Pi^\perp_{\widehat{T}_{\theta_t}\mathcal{M}}\nabla\mathcal{L}_{|u_{\theta_t}}\right) \simeq \widehat{\phi}^\dagger_{\theta_t}\widehat{\nabla\mathcal{L}}_{\theta_t},$$

with: for all $1 \leqslant p \leqslant P, 1 \leqslant i \leqslant S$

- $\widehat{\phi}_{\theta_t\, i,p} := \partial_p u_{\theta_t}(x_i)$
- $\widehat{\nabla\mathcal{L}}_{\theta_t\, i} := \nabla\mathcal{L}_{|u_{\theta_t}}(x_i)$

Key fact

$\hat{\phi}^\dagger_{\theta_t}$ **can be computed with a SVD, with complexity** $\mathrm{O}\big(\min(\mathrm{PS}^2, \mathrm{P}^2\mathrm{S})\big)$.

Corollary

There exists P points $(\hat{x}_i)$ *such that:*

$$\Pi^\perp_{\widehat{T}_\theta\mathcal{M}}\nabla\mathcal{L}_{|u_\theta} = \Pi^\perp_{T_\theta\mathcal{M}}\nabla\mathcal{L}_{|u_\theta}.$$

Application to PINNs

Key remark

The only difference between the losses:

$$\widehat{\ell}_{D,B}(\theta) := \frac{1}{2S_D}\sum_{i=1}^{S_D}\left(D[u_\theta](x_i^D) - f(x_i^D)\right)^2 + \frac{1}{2S_B}\sum_{i=1}^{S_B}\left(B[u_\theta](x_i^B) - g(x_i^B)\right)^2,$$

and $\widehat{\ell}(u) := \frac{1}{2S}\sum_{i=1}^{S}\left(u(x_i) - f(x_i)\right)^2$ is the use of the operators D and B.

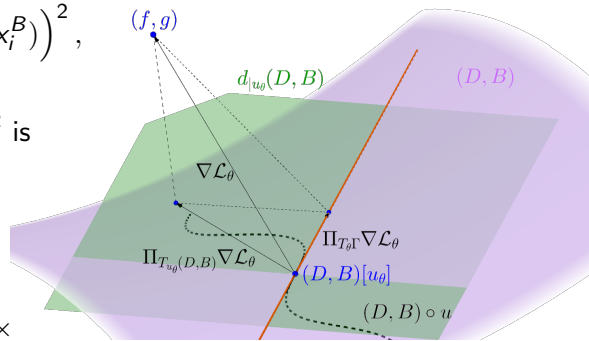
Proposition

PINNs are a quadradic regression problem with model: $(D, B) \circ u$:

$$\begin{cases} \mathbb{R}^P & \to & \mathcal{H} & \to & L^2(\Omega \to \mathbb{R}, \mu) \times \\ & & & & L^2(\partial\Omega \to \mathbb{R}, \sigma) \\ \theta & \mapsto & u_\theta & \mapsto & (D[u_\theta], B[u_\theta]) \end{cases}$$

Natural Gradient of PINNs

Figure: Illustration of PINNs Natural Gradient



Empirical Evidence for the Natural Gradient Relative to Adapted Model (*ANaGRAM*) Algorithm

2 D Laplace equation

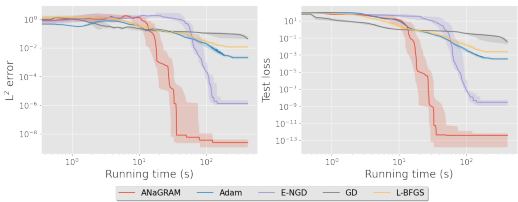


Figure: Performance comparison w.r.t running time for Laplace equation in 2 D:

$$\begin{cases} \Delta u = -2\pi^2 \sin(\pi x_1)\sin(\pi x_2) & \text{in } [0,1]^2 \\ u = 0 & \text{on } \partial[0,1]^2 \end{cases}$$

1+1 D Heat equation

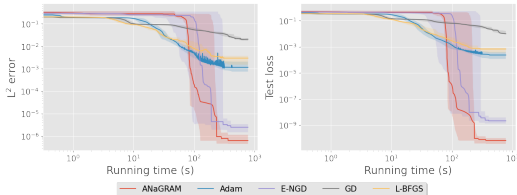


Figure: Performance comparison w.r.t running time for Heat equation in 1+1 D:

$$\begin{cases} \partial_t u - \frac{1}{4}\partial_{xx} u = 0 & \text{in } [0,1]^2 \\ u = 0 & \text{on } [0,1] \times \{0,1\} \\ u = sin(\pi x) & \text{on } \{0\} \times [0,1] \end{cases}$$

5 D Laplace equation

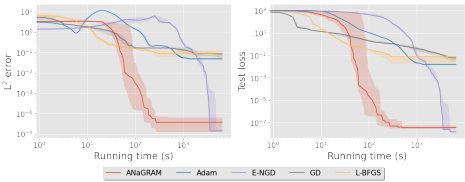


Figure: Performance comparison w.r.t running time for Laplace equation in 5 D:

$$\begin{cases} \Delta u = \pi^2 \sum_{k=1}^{5} \sin(\pi x_k) & \text{in } \Omega = [0,1]^5 \\ u = \sum_{k=1}^{5} \sin(\pi x_k) & \text{on } \partial\Omega \end{cases}$$

1+1 D Allen-Cahn equation

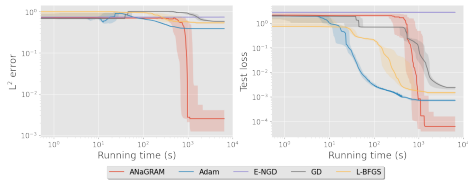
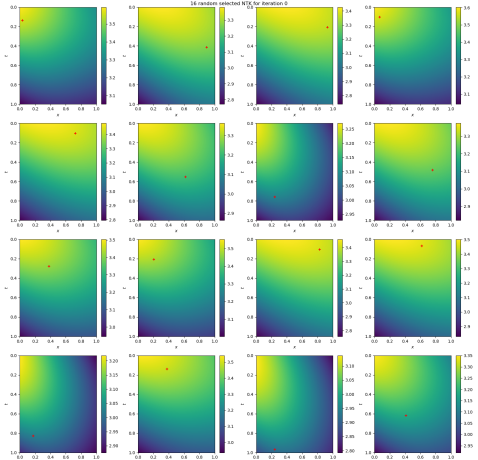
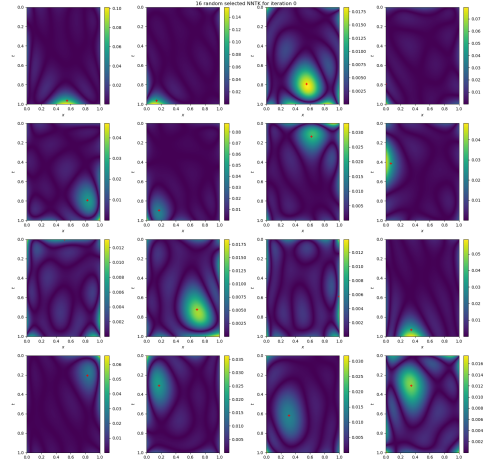


Figure: Performance comparison w.r.t running time for Allen-Cahn equation in 1+1 D:

$$\begin{cases} \partial_t u - 10^{-3}\, \partial_{xx} u = 5(u - u^3) & \text{in } \Omega = [0,1] \times [-1,1] \\ u = -1 & \text{on } \partial\Omega_b = [0,1] \times \{-1,1\} \\ u(0,x) = x^2 \cos(\pi x) & \text{on } \partial\Omega_0 = \{0\} \times [-1,1] \end{cases}$$

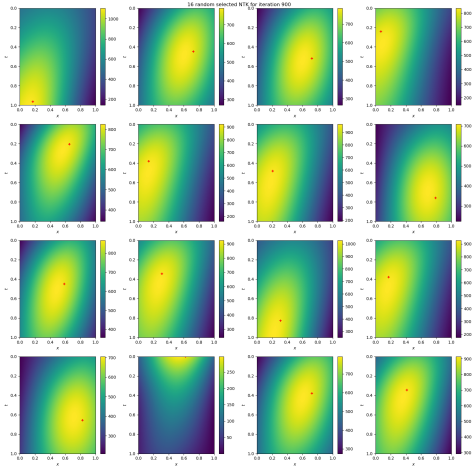


(a) NTK

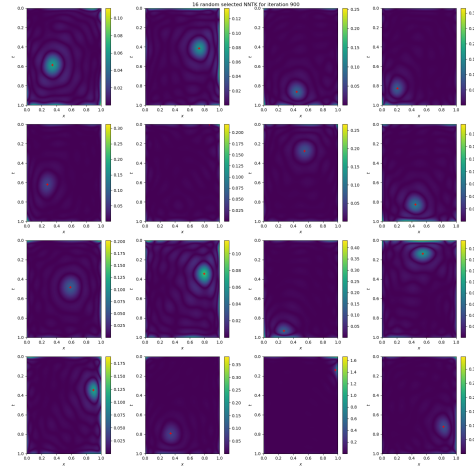


(b) NNTK

Figure: Comparison of NTK and NNTK at **initialization** for Heat equation in 1+1 D



(a) NTK



(b) NNTK

Figure: Comparison of NTK and NNTK at the **end of optimization** for Heat equation in 1+1 D

In-Depth Empirical Analysis of *Cutoff* Regularization in *ANaGRAM*

In-Depth Empirical Analysis of *Cutoff* Regularization in *ANaGRAM*

SVD pseudoinverse details

$$\widehat{\phi}_\theta = \widehat{V}\,\widehat{\Delta}\,\widehat{U}^\top; \qquad \widehat{\phi}^\dagger_{\theta_t} = \widehat{U}\widehat{\Delta}^\dagger\widehat{V}^\top.$$

Singular values

In practice, we apply a *cutoff*:

$$\widehat{\Delta}^{\dagger_\alpha} := \begin{cases} \widehat{\Delta}_i^{-1} & \text{if } \widehat{\Delta}_i \geqslant \alpha \\ 0 & \text{otherwise} \end{cases},$$

with $\alpha > 0$ the cutoff level. Thus:

$$\widehat{\phi}^{\dagger_\alpha}_\theta \widehat{\nabla\mathcal{L}}_\theta = \sum_{i=1}^{r_\alpha} \widehat{U}_i \widehat{\Delta}_i^{-1} \widehat{V}_i^\top \widehat{\nabla\mathcal{L}}_\theta,$$

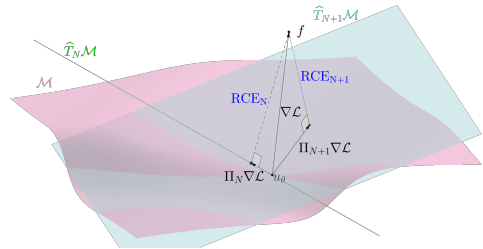
with $r_\alpha := \#\{i : \widehat{\Delta}_i \geqslant \alpha\} \leqslant \min(P, S)$.

Reconstruction Error (RCE)

$$\mathrm{RCE}_n = \frac{1}{\sqrt{S}} \left\| \widehat{\nabla\mathcal{L}}_\theta - \sum_{i=1}^{n} \widehat{V}_i \widehat{V}_i^\top \widehat{\nabla\mathcal{L}}_\theta \right\|_{\mathbb{R}^S}$$

Intuition on RCE

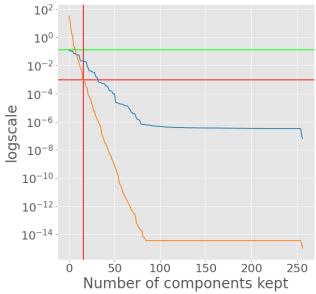
RCE_n accounts for the part of $\nabla\mathcal{L}_\theta$ orthogonal to the n most important components of $\widehat{T}_\theta\mathcal{M}$.



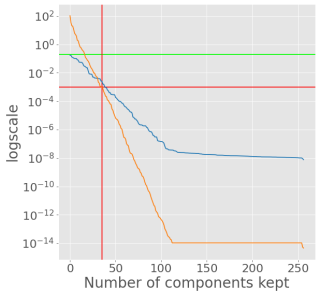
Remark

- $\lim_{S\to\infty} \widehat{T}_\theta\mathcal{M} = T_\theta\mathcal{M}$
- $\mathrm{RCE}_0^2 = \frac{1}{S}\left\|\widehat{\nabla\mathcal{L}_\theta}\right\|^2_{\mathbb{R}^S} = \widehat{\ell}(\theta)$

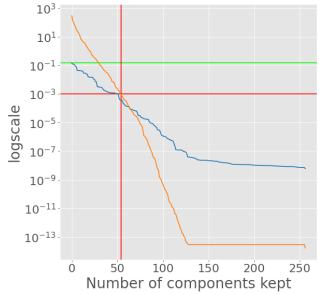
Empirical insights on the *cutoff* impact in *ANaGRAM*



(a) Iteration 0: intersection point between singular values and RCE lies before cutoff.

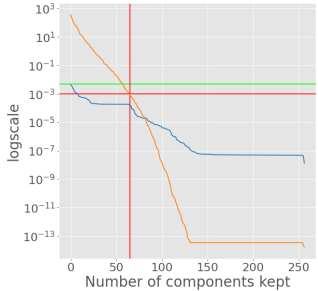


(b) Iteration 40: intersection point shifts rightward toward cutoff.

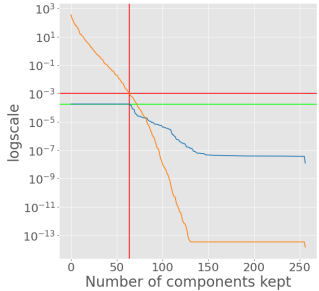


(c) Iteration 90: intersection point passes the cutoff threshold.

Empirical insights on the *cutoff* impact in *ANaGRAM*

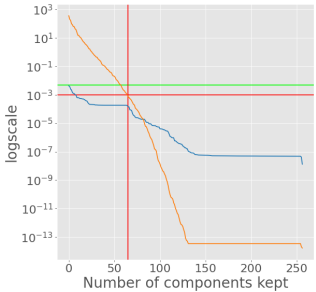


(d) Iteration 120. Beginning of *flattening*: RCE stabilizes at constant level before cutoff.

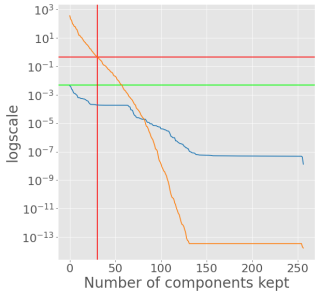


(e) Iteration 150: End of the *flattening phenomenon*. The train loss reaches the flattened part of the RCE.

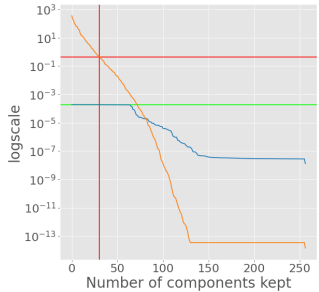
Incomplete and instant *flattening*



(f) Incomplete flattening of the RCE with a fixed cutoff at 10^{-3}.



(g) New cutoff located roughly at the location of the "elbow" in the RCE curve.



(h) Complete flattening after one natural gradient step with the new cutoff.

Adaptive Multi-cutoff Strategy Modification for ANaGRAM (*AMStramGRAM*) algorithm

Algorithm 1: *AMStramGRAM* (sketch)

Input:

Initial parameters: $\theta_0 \in \mathbb{R}^P$

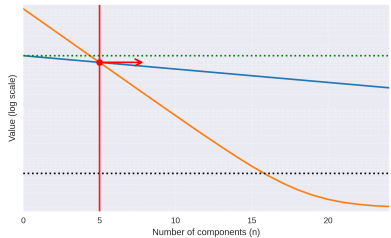
Precision target: $\epsilon > 0$

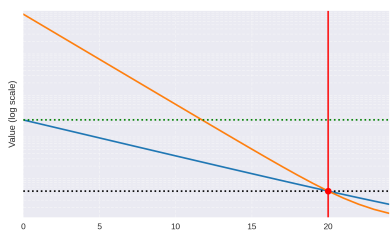
repeat

- $\widehat{U}_t, \widehat{\Delta}_t, \widehat{V}_t \leftarrow \mathrm{SVD}(\widehat{\phi}_t)$
- Compute RCE_t
- $r_\cap \leftarrow \#\{n : RCE_{t_n} \geqslant \widehat{\Delta}_{t_n}\}$
- $r_\epsilon \leftarrow \#\{n : RCE_{t_n} \geqslant \epsilon\}$
- Apply ANaGRAM with cutoff rank $r_\alpha \leftarrow \min(r_\cap, r_\epsilon)$
- $t \leftarrow t + 1$

until $r_\epsilon = 0$ *or* $t \geqslant T_{\max}$

Output: θ_t





Algorithm 1: *AMStramGRAM* (sketch)

Input:

Initial parameters: $\theta_0 \in \mathbb{R}^P$

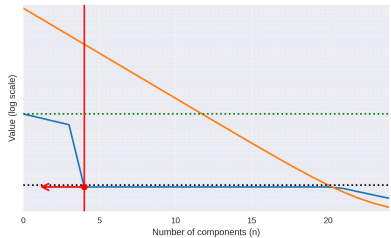
Precision target: $\epsilon > 0$

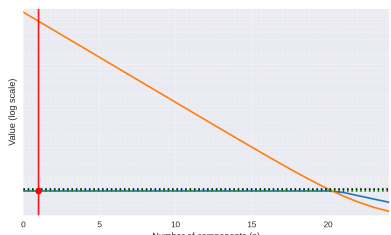
repeat

- $\widehat{U}_t, \widehat{\Delta}_t, \widehat{V}_t \leftarrow \mathrm{SVD}(\widehat{\phi}_t)$
- Compute RCE_t
- $r_\cap \leftarrow \#\{n : RCE_{t_n} \geqslant \widehat{\Delta}_{t_n}\}$
- $r_\epsilon \leftarrow \#\{n : RCE_{t_n} \geqslant \epsilon\}$
- Apply ANaGRAM with cutoff rank $r_\alpha \leftarrow \min(r_\cap, r_\epsilon)$
- $t \leftarrow t + 1$

until $r_\epsilon = 0$ *or* $t \geqslant T_{\max}$

Output: θ_t





Experiment	Train Loss		L_2 Error	
	AMStraMGRAM	ANaGRAM	AMStraMGRAM	ANaGRAM
Heat Equation	**6.29e-29 ± 6.78e-30**	8.56e-11 ± 7.05e-11	**2.32e-14 ± 1.14e-14**	1.28e-06 ± 1.75e-06
Laplace 2D	**1.46e-28 ± 1.87e-29**	4.27e-13 ± 4.66e-13	**2.24e-15 ± 2.52e-16**	3.49e-09 ± 3.58e-09
Laplace 5D	**2.04e-08 ± 1.16e-08**	6.37e-08 ± 7.01e-08	**2.12e-05 ± 8.15e-06**	4.00e-05 ± 2.93e-05
Allen–Cahn	**3.19e-11 ± 2.37e-11**	2.19e-04 ± 4.16e-04	**5.87e-05 ± 6.25e-06**	4.32e-03 ± 5.93e-03

Experiment	Train Loss		L_2 Error	
	AMStraMGRAM	SSBroyden	AMStraMGRAM	SSBroyden
Burgers (1+1 D)	**2.99e-12 ± 9.26e-13**	2.92e-10 ± 1.45e-10	**1.5e-06 ± 9.43e-7**	1.59e-06 ± 1.02e-6
Non-Linear Poisson	**8.51e-24 ± 2.24e-24**	3.03e-16 ± 3.82e-16	6.81e-10 ± 1.41e-09	**9.29e-12 ± 5.85e-12**
Allen–Cahn	3.19e-11 ± 2.37e-11	**6.42e-12 ± 5.52e-12**	5.87e-05 ± 6.25e-06	**3.94e-06 ± 1.72e-06**

Overfitting

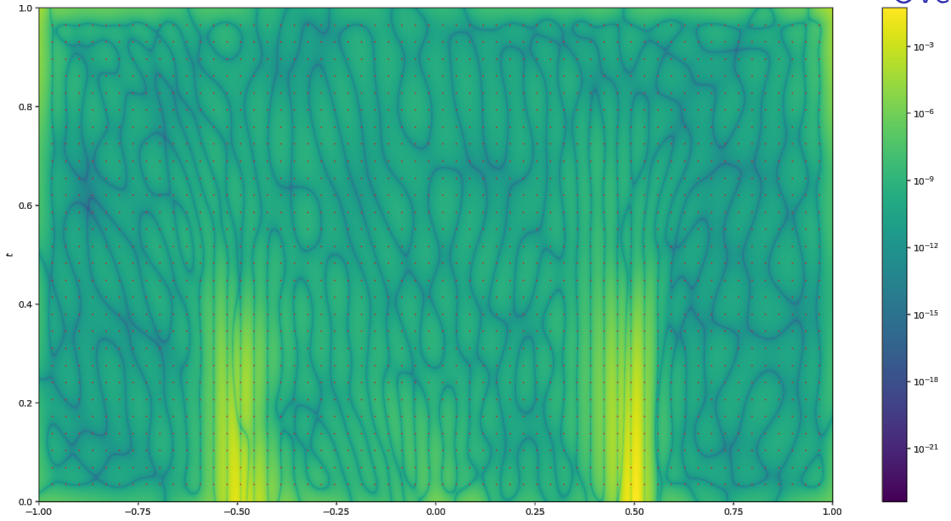


Figure: Overfitting on Allen–Cahn: residual lines align with sampling lines.

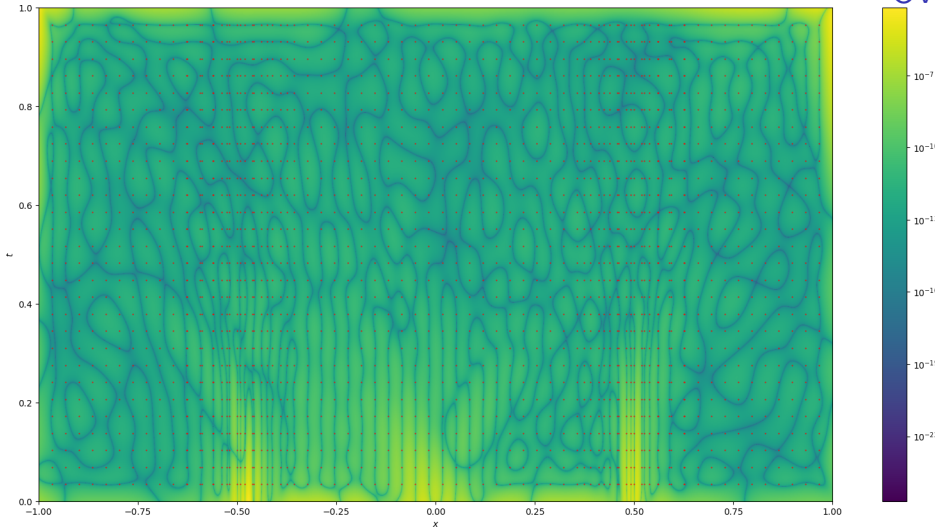


Figure: Overfitting on Allen–Cahn: densifying the sampling in overfitted regions mitigates overfitting.

Natural Gradient and Green's function

Definition (Green's function of D)

A Green's function is any kernel function $g : \Omega \times \Omega \to \mathbb{R}$ such that the operator:

$$R : f \in D\left[\mathcal{H}\right] \mapsto \left(x \in \Omega \mapsto \int_{\Omega} g(x,s) f(s) \mu(\mathrm{d}s) \right) \in \mathcal{H}$$

verifies the equation: $D \circ R = I_{D[\mathcal{H}]}$

Definition (generalized Green's function of D on $\mathcal{H}_0 \subset \mathcal{H}$)

A generalized Green's function is any kernel function $g : \Omega \times \Omega \to \mathbb{R}$ such that the operator:

$$R : f \in \mathrm{L}^2(\Omega \to \mathbb{R}, \mu) \mapsto \left(x \in \Omega \mapsto \int_{\Omega} g(x,s) f(s) \mu(\mathrm{d}s) \right) \in \mathcal{H}$$

verifies the equation: $D \circ R = \Pi^{\perp}_{D[\mathcal{H}_0]}$

Theorem

Let $D : \mathcal{H} \to L^2(\Omega \to \mathbb{R}, \mu)$ be a linear differential operator and $u : \mathbb{R}^P \to \mathcal{H}$ a parametric model. Then $for\ all\ \theta \in \mathbb{R}^P$, the generalized Green's function of D on $T_\theta\mathcal{M} = \mathrm{Im}\, du_\theta$ is given by: $for\ all\ x, y \in \Omega$

$$g_{T_\theta\mathcal{M}}(x, y) := \sum_{1 \leqslant p,q \leqslant P} \partial_p u_\theta(x)\, G^\dagger_{p,q} \partial_q D[u_\theta](y),$$

with: $for\ all\ 1 \leqslant p, q \leqslant P$

$$G_{pq} := \langle \partial_p D[u_\theta]\,,\, \partial_q D[u_\theta] \rangle_{L^2(\Omega \to \mathbb{R}, \mu)}\,.$$

In particular, the natural gradient of PINNs can be rewritten:

$$\theta_{t+1} \leftarrow \theta_t - \eta\, du^\dagger_{\theta_t} \left(x \in \Omega \mapsto \int_\Omega g_{T_{\theta_t}\mathcal{M}}(x, y) \nabla \mathcal{L}_{\theta_t}(y) \mu(dy) \right).$$

Theorem

Let $D : \mathcal{H} \to L^2(\Omega \to \mathbb{R}, \mu)$ be a linear differential operator and $\mathcal{H}_0 \subset \mathcal{H}$ an RKHS with kernel k_0. Given the spectral decomposition:

$$\Pi_{\mathcal{H}_0} D^* D \Pi_{\mathcal{H}_0} = \int_0^{+\infty} \lambda \pi_{D,\mathcal{H}_0}(d\lambda) \qquad I_{\mathcal{H}_0} = \int_0^{+\infty} \pi_{D,\mathcal{H}_0}(d\lambda)$$

Then the generalized Green's function at regularization level $\alpha > 0$ is given by:
for all $x, y \in \Omega$

$$g_{\mathcal{H}_0,\alpha}(x, y) := D \left[\int_{\alpha^2}^{+\infty} \lambda^{-1} \pi_{D,\mathcal{H}_0}(d\lambda) \left[k(x, \cdot) \right] \right] (y)$$

Conclusion and Perspectives

Conclusions

- Anagram: lowers natural-gradient cost from $O(P^3)$ to $O(\min(PN^2, P^2N))$ (overfactor $\min(P, N)$ w.r.t SGD).
- AMStramGRAM gives a principled way to adapt cutoff reaching machine-level error.
- We prove that PINNs natural gradient corresponds to an optimal linear update following the Green's function.
- Empirical results are competitive with state-of-the-art PINNs optimizers.

Perspectives

- Design of an optimal collocation points procedure, coupled with AMStramGRAM's cutoff adaptation strategy.
- Include data assmilation in this theoretical setting, and understand its regularizing effect.
- Include common optimization techniques (*e.g.* Momentum)
- Extend it to Operator learning

Thank you for your attention !

Publications

- Schwencke, N. and C. Furtlehner (2025): "ANaGRAM: A Natural Gradient Relative to Adapted Model for Efficient PINNs Learning," in The Thirteenth International Conference on Learning Representations.
- Marie-Anne, J., C. Rousselot, N. Schwencke, and A. Shilova (2025): "Implicit Function Theorem in Physics-Informed Neural Networks to Solve Parameterized Differential Equations," in EurIPS 2025 Workshop: Differentiable Systems and Scientific Machine Learning.

Preprint

- Schwencke, N., C. Rousselot, A. Shilova, and C. Furtlehner (2025): "AMStramGRAM: Adaptive Multi-Cutoff Strategy Modification for ANaGRAM," arXiv Preprint.

References I

Amari, S.-I. and S. C. Douglas (1998): "Why Natural Gradient?" in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, IEEE, vol. 2, 1213–1216.

Berezansky, Y. M., Z. G. Sheftel, and G. F. Us (1996): *Functional Analysis. Vol. II*, vol. 86 of *Operator Theory Advances and Applications*, Birkhäuser.

Brunton, S. L., J. L. Proctor, and J. N. Kutz (2016): "Discovering Governing Equations from Data by Sparse Identification of Nonlinear Dynamical Systems," *Proceedings of the National Academy of Sciences*, 113, 3932–3937.

Chen, R. T., Y. Rubanova, J. Bettencourt, and D. K. Duvenaud (2018): "Neural Ordinary Differential Equations," *Advances in neural information processing systems*, 31.

Jacot, A., F. Gabriel, and C. Hongler (2018): "Neural Tangent Kernel: Convergence and Generalization in Neural Networks," *Advances in neural information processing systems*, 31.

Lagaris, I. E., A. Likas, and D. I. Fotiadis (1998): "Artificial Neural Networks for Solving Ordinary and Partial Differential Equations," *IEEE transactions on neural networks*, 9, 987–1000.

Li, Z., N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, and A. Anandkumar (2020): "Fourier Neural Operator for Parametric Partial Differential Equations," *arXiv preprint arXiv:2010.08895*.

Lu, L., P. Jin, and G. E. Karniadakis (2021): "DeepONet: Learning Nonlinear Operators for Identifying Differential Equations Based on the Universal Approximation Theorem of Operators," *Nature Machine Intelligence*, 3, 218–229.

Raissi, M., P. Perdikaris, and G. Karniadakis (2019): "Physics-Informed Neural Networks: A Deep Learning Framework for Solving Forward and Inverse Problems Involving Nonlinear Partial Differential Equations," *Journal of Computational Physics*, 378, 686–707.

Rudner, T. G., F. Wenzel, Y. W. Teh, and Y. Gal (2019): "The Natural Neural Tangent Kernel: Neural Network Training Dynamics under Natural Gradient Descent," in *4th Workshop on Bayesian Deep Learning (NeurIPS 2019)*.

Urbán, J. F., P. Stefanou, and J. A. Pons (2025): "Unveiling the Optimization Process of Physics Informed Neural Networks: How Accurate and Competitive Can PINNs Be?" *Journal of Computational Physics*, 523, 113656.

Ongoing work : Implicit Curriculum Learning

- Marie-Anne, J., C. Rousselot, N. Schwencke, and A. Shilova (2025): "Implicit Function Theorem in Physics-Informed Neural Networks to Solve Parameterized Differential Equations," in EurIPS 2025 Workshop: Differentiable Systems and Scientific Machine Learning.

Implicit Curriculum Learning

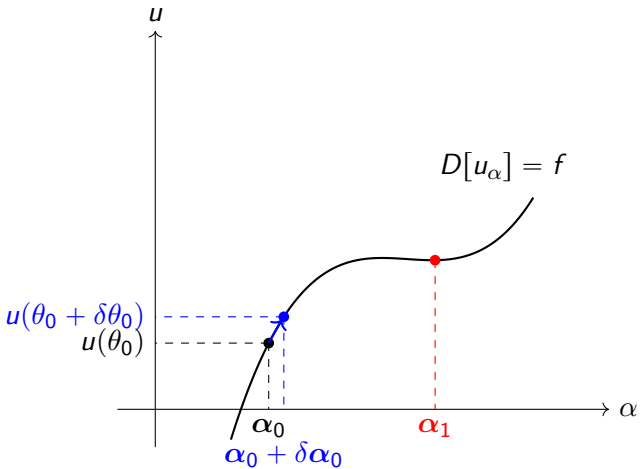


Figure: Illustration of implicit function theorem.

Credits : Julien Marie-Anne

Parameterized PDE

Let $\alpha \in \mathcal{A} \subset \mathbb{R}$ and consider

$$D[\alpha, u] = f \quad \text{in } \Omega.$$

Implicit function theorem

If $\mathrm{d}D_{\alpha,u} \neq 0$, there is $[a, b] \ni \alpha$ and v such that for all $\beta \in [a, b]$

$$D[\beta, v(\beta)] = f.$$

with v solution to the ODE:

$$\frac{\mathrm{d}}{\mathrm{d}\beta} v = \partial_u D^{-1}_{\beta, v(\beta)} \left[\frac{\mathrm{d}}{\mathrm{d}\beta} D_{\beta, v(\beta)} \right]$$

Implicit curriculum learning

Follow the update:

$$\frac{\mathrm{d}}{\mathrm{d}\beta} \theta(\beta) = \partial_\theta (D \circ u_\theta)^\dagger_{\beta, u_\theta(\beta)} \left[\frac{\mathrm{d}}{\mathrm{d}\beta} D_{\beta, u_\theta(\beta)} \right]$$

Table: Comparison of the method with Adam on different PDEs

Equation	**Method**	**Metric**
Hamilton–Jacobi–Bellman (Relative Error)	Adam	5.56e-01 $\pm$ 1.99e-01
	Implicit Curriculum Learning	**3.44e-01** $\pm$ **2.05e-01**
Eikonal (Relative Error)	Adam	8.02e-01 $\pm$ 9.78e-01
	Implicit Curriculum Learning	**2.05e-02** $\pm$ **1.09e-02**
Burgers (Evaluation Loss)	Adam	1.32e-02
	Implicit Curriculum Learning	**7.21e-03**

Ongoing work : *Collocation points* selection

Remark

$$(x_i)^\star = \underset{(x_i)\in\Omega^S}{\operatorname{argmin}} \left\| \Pi^{\perp}_{\widehat{T}^{(x_i)}_{\theta,K}\mathcal{M}} \nabla\mathcal{L}_{|u_{\theta_t}} - \nabla\mathcal{L}_{|u_{\theta_t}} \right\|_{\mathrm{L}^2} = \underset{(x_i)\in\Omega^S}{\operatorname{argmin}} \inf_{\alpha\in\mathbb{R}^S} \left\| \sum_{i=1}^{S} \alpha_i K_\theta(x_i,\cdot) - \nabla\mathcal{L}_{|u_{\theta_t}} \right\|^2_{\mathrm{L}^2}$$

Consequence

$(x_i)^\star$ can be "learned" by the minimization through natural gradient descent of

$$u : \begin{cases} \Omega^S \times \mathbb{R}^S & \to & \mathrm{L}^2(\Omega \to \mathbb{R}, \mu) \\ \big((x_i),\alpha\big) & \mapsto & \sum_{i=1}^{S} \alpha_i K_\theta(x_i,\cdot) \end{cases}$$

Even better: closed form formulas exist !

Proposition

- $\left\langle \partial_{\alpha_i} u_\theta, \partial_{\alpha_j} u_\theta \right\rangle = K_\theta(x_i,x_j)$
- $\left\langle \partial_{x_i} u_\theta, \partial_{\alpha_j} u_\theta \right\rangle = \alpha_i \partial_1 K_\theta(x_i,x_j)$
- $\left\langle \partial_{x_i} u_\theta, \partial_{x_j} u_\theta \right\rangle = \alpha_i \partial_2\partial_1 K_\theta(x_i,x_j)\alpha_j.$
- $\left\langle \partial_{\alpha_i} u_\theta, \nabla\mathcal{L} \right\rangle = \Pi_{T_\theta\mathcal{M}} \nabla\mathcal{L}(x_i) \simeq \nabla\mathcal{L}(x_i)$
- $\left\langle \partial_{x_i} u_\theta, \nabla\mathcal{L} \right\rangle = \alpha_i \Pi_{T_\theta\mathcal{M}} \nabla\mathcal{L}'(x_i) \simeq \alpha_i \nabla\mathcal{L}'(x_i)$

First results for collocation learning in Fourier space

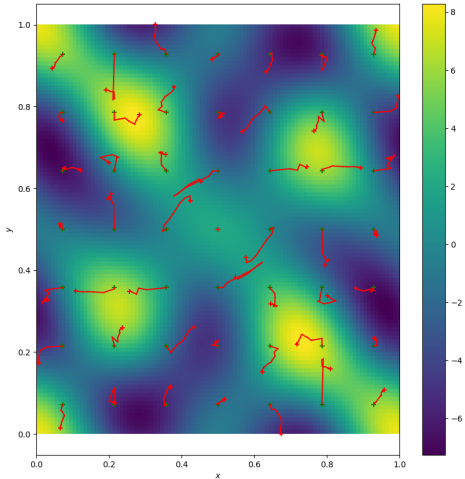


Figure: Points learning dynamic

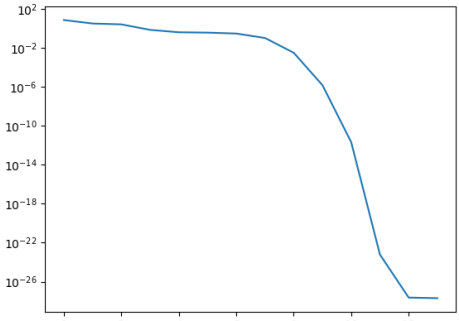


Figure: $\left\| \Pi^{\perp}_{\widehat{T}^{(x_j)}_{\theta,K}\mathcal{M}} \nabla \mathcal{L}_{|u_{\theta_t}} - \nabla \mathcal{L}_{|u_{\theta_t}} \right\|_{\mathbf{L^2}}$ wrt (x_i) learning steps

Ongoing work : Connection to Galerkin's method

Strong formulation

Find $u \in \mathrm{H}^2(\Omega)$ such that

$$\begin{cases} \Delta u = f \in \mathrm{L}^2(\Omega) & \text{in } \Omega \\ u = 0 \in \mathrm{L}^2(\partial\Omega) & \text{on } \partial\Omega \end{cases}.$$

Weak formulation

Find $u \in \mathrm{H}_0^1(\Omega)$ such that:$\forall v \in \mathrm{H}_0^1(\Omega)$,

$$\langle \nabla u\,,\, \nabla v \rangle_{\mathrm{L}^2(\Omega\to\mathbb{R}^d)} = \langle v\,,\, -f \rangle_{\mathrm{L}^2(\Omega)}.$$

Galerkin method

Fixing a finite dimensional space $\mathrm{H}_n \subset \mathrm{H}_0^1(\Omega)$.

Find $u \in \mathrm{H}_n$ such that:$\forall v \in \mathrm{H}_n$,

$$\langle \nabla u\,,\, \nabla v \rangle_{\mathrm{L}^2(\Omega\to\mathbb{R}^d)} = \langle v\,,\, -f \rangle_{\mathrm{L}^2(\Omega)}.$$

Kernelization of Galerkin's Method

Kernelization of Galerkin

Let $(v_i^n)_{i=1}^n$ be a basis of H_n.

$$\mathcal{T}_{\mathrm{H}_n} : \theta \in \mathbb{R}^n \mapsto \sum_{i=1}^{n} \theta_i v_i^n \in \mathrm{H}_n.$$

$$\mathrm{NNTK_s(x,y)} = \sum_{1\leqslant \mathrm{p,q} \leqslant \mathrm{n}} \partial_{\mathrm{p}} \mathcal{T}_{\mathrm{H_n}} \mathrm{G}^{\dagger}_{\mathrm{s_{p,q}}} \partial_{\mathrm{q}} \mathcal{T}_{\mathrm{H_n}}.$$

Strong formulation

$$G2_{p,q} = \langle \Delta\partial_p \mathcal{T}_{\mathrm{H}_n}\,,\, \Delta\partial_q \mathcal{T}_{\mathrm{H}_n} \rangle_{\mathrm{L}^2(\Omega)}.$$

$$u_n = \langle \Delta \mathrm{NNTK_2(x,\cdot)}\,,\, \mathrm{f} \rangle_{\mathrm{L}^2(\Omega)}.$$

Weak formulation

$$G1_{p,q} = \langle \nabla\partial_p \mathcal{T}_{\mathrm{H}_n}\,,\, \nabla\partial_q \mathcal{T}_{\mathrm{H}_n} \rangle_{\mathrm{L}^2(\Omega\to\mathbb{R}^d)}.$$

$$u_n = \langle \mathrm{NNTK_1(x,\cdot)}\,,\, -\mathrm{f} \rangle_{\mathrm{L}^2(\Omega)}.$$

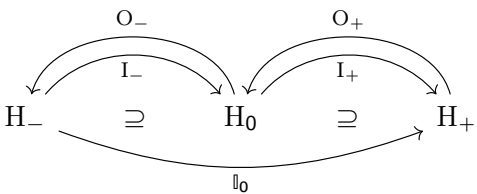


Figure: Schematic diagram of the basic structure of a Hilbert Rigging (a.k.a. Gelfand triple), adapted from Berezansky et al. (1996).

$$\mathrm{O}_+ : \text{embedding of } \mathrm{H}_+ \text{ into } \mathrm{H}_0$$

$$\mathrm{I}_+ := \mathrm{O}_+^* : \mathrm{H}_0 \to \mathrm{H}_+$$

$$\langle \cdot\,, \cdot \rangle_- : (u, v) \in \mathrm{H}_0^2 \mapsto \langle \mathrm{I}_+ u\,, \mathrm{I}_+ v \rangle_{\mathrm{H}_+}$$

$$\mathrm{H}_- := \overline{\mathrm{H}_0}^{\|\cdot\|_-} \simeq \mathrm{H}_+'$$

$$\mathbb{I}_0 := \overline{\mathrm{I}_+} : \mathrm{H}_- \to \mathrm{H}_+$$

$$\mathrm{O}_- : \text{embedding of } \mathrm{H}_0 \text{ into } \mathrm{H}_-$$

$$\mathrm{I}_- := \mathrm{O}_-^* : \mathrm{H}_- \to \mathrm{H}_0$$

$$\mathbb{I}_0 \mathrm{O}_- = \mathrm{I}_+; \quad \mathrm{O}_+ \mathbb{I}_0 = \mathrm{I}_-$$

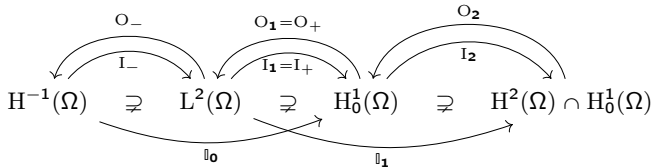


Figure: Schematic representation of the Hilbert rigging chain unifying the weak and strong formulations. Adapted from Berezansky et al. (1996).

Proposition (Green's Identity)

$$\langle \Delta u\,,\, O_1 v\rangle_{L^2(\Omega)} = -\int_\Omega \langle \nabla u\,,\, \nabla v\rangle_{\mathbb{R}^d} = -\langle O_2 u\,,\, v\rangle_{H^1_0(\Omega)}\,.$$

Proposition

$$\Delta_2 := \Delta = -\mathbb{I}_1^* : H^2(\Omega)\cap H^1_0(\Omega) \to L^2(\Omega)$$

$$\Delta_1 := \overline{\Delta}^{H^1_0(\Omega)} = -\mathbb{I}_0^* : H^1_0(\Omega) \to H^{-1}(\Omega)$$

Weak-Strong equivalence

Test form of the Strong formulation

Find $u \in \mathrm{H}^2(\Omega)\cap \mathrm{H}^1_0(\Omega)$ s.t,

$\langle \Delta u\,,\, v\rangle_{\mathrm{H}^1_0(\Omega)} = -\langle f\,,\, v\rangle_{\mathrm{H}^1_0(\Omega)}\,,$

$\forall v \in \mathrm{H}^2(\Omega)\cap \mathrm{H}^1_0(\Omega).$

Least-squares form of the Weak formulation

Find $u \in \mathrm{H}^1_0(\Omega)$ s.t $\Delta_1[u] = \mathrm{O}_- f$

or even $\Delta_1[u] = \alpha \in \mathrm{H}^{-1}(\Omega)$.

Proposition

$$u_{\text{weak}} = -I_1 f = O_2 u_{strong}$$

Green's functions

Form \ Formulation	Weak formulation	Strong formulation
Test form	$-\mathrm{O}_1[\mathrm{NNTK}_{1,\mathrm{n}}(\mathrm{x},\cdot)] \simeq -\mathrm{NNTK}_{1,\mathrm{n}}(\mathrm{x},\cdot) \in \mathrm{L}^2(\Omega)$	$-\mathrm{O}_2[\mathrm{NNTK}_{2,\mathrm{n}}(\mathrm{x},\cdot)] \simeq -\mathrm{NNTK}_{2,\mathrm{n}}(\mathrm{x},\cdot) \in \mathrm{H}_0^1(\Omega)$
Least-squares form	$-\mathbb{I}_0^*[\mathrm{NNTK}_{1,\mathrm{n}}(\mathrm{x},\cdot)] = \Delta_1[\mathrm{NNTK}_{1,\mathrm{n}}(\mathrm{x},\cdot)] \in \mathrm{H}^{-1}(\Omega)$	$-\mathbb{I}_1^*[\mathrm{NNTK}_{2,\mathrm{n}}(\mathrm{x},\cdot)] = \Delta_2[\mathrm{NNTK}_{2,\mathrm{n}}(\mathrm{x},\cdot)] \in \mathrm{L}^2(\Omega)$

Table: Summary of the Green's functions associated with each combination of formulation and form, computed over a finite-dimensional subspace H_n of functions, with $x \in \Omega$.

N.B: While those approximations yield the same asymptotic solution, the approximations (*i.e.* solutions on H_n) differ.

Illustration of natural gradient dynamics

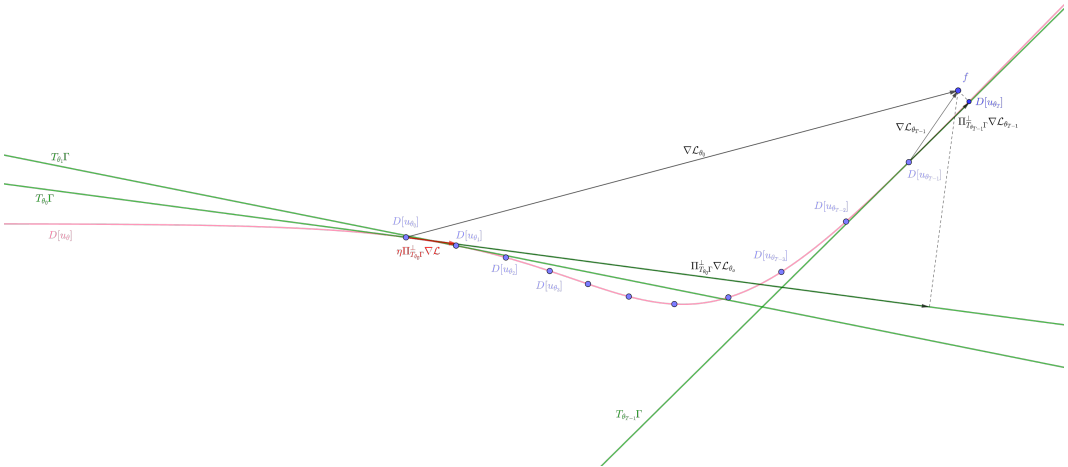


Figure: Illustration of PINNs learning process under natural gradient, as successive applications of Green's function

In the population limit, the natural gradient of PINNs is the update:

$$\theta_{t+1} \leftarrow \theta_t - \eta \, \mathrm{d}\big((D,B)\circ u\big)^{\dagger}_{\theta_t} \left(\Pi^{\perp}_{T_{\theta_t}\Gamma} \nabla \mathcal{L}_{|u_{\theta_t}}\right)$$

Corollary

The kernel of $\Pi_{T_\theta \Gamma}$ *is: for all* $x, y \in (\Omega \times \partial\Omega)^2$

$$\begin{aligned} NNTK_\theta(x,y) &= \sum_{1\leqslant p,q\leqslant P} \partial_p(D,B)\left[u_\theta\right](x) G_{\theta\,p,q}^{\dagger} \partial_q(D,B)\left[u_\theta\right](y) \\ &= \sum_{1\leqslant p,q\leqslant P} \left(\partial_p D\left[u_\theta\right](x_1), \partial_p B\left[u_\theta\right](x_2)\right) G_{\theta\,p,q}^{\dagger} \left(\partial_q D\left[u_\theta\right](y_1), \partial_q B\left[u_\theta\right](y_2)\right), \end{aligned}$$

where $for\ all\ 1 \leqslant p, q \leqslant P$

$$\begin{aligned} G_{\theta\,p,q} &:= \left\langle \partial_p(D,B)\left[u_\theta\right] \,,\, \partial_q(D,B)\left[u_\theta\right] \right\rangle_{L^2(\Omega\to\mathbb{R},\mu)\times L^2(\partial\Omega\to\mathbb{R},\sigma)} \\ &= \left\langle \partial_p D\left[u_\theta\right] \,,\, \partial_q D\left[u_\theta\right] \right\rangle_{L^2(\Omega\to\mathbb{R},\mu)} + \left\langle \partial_p B\left[u_\theta\right] \,,\, \partial_q B\left[u_\theta\right] \right\rangle_{L^2(\partial\Omega\to\mathbb{R},\sigma)}. \end{aligned}$$

Corollary

The kernel of $\Pi_{\widehat{T}_\theta \Gamma}$ *is: for all* $x, y \in (\Omega \times \partial\Omega)^2$

$$\widehat{k}(x,y) = \sum_{1 \leqslant i,j \leqslant S} NNTK_\theta(x, x_i)\, \widehat{G}^{\dagger}_{\theta\, i,j}\, NNTK_\theta(x_j, y), \text{ where}$$

$$G_{\theta\, i,j} := \left\langle NNTK_\theta(\cdot, x_i)\,,\, NNTK_\theta(x_j, \cdot) \right\rangle_{L^2(\Omega \to \mathbb{R}, \mu) \times L^2(\partial\Omega \to \mathbb{R}, \sigma)} = NNTK_\theta(x_i, x_j)$$

Theorem (ANaGRAM for PINNs)

Under mild assumptions, the empirical natural gradient update:

$$\theta_{t+1} \leftarrow \theta_t - \eta\, d\big((D,B) \circ u\big)^{\dagger}_{\theta_t} \left(\Pi^{\perp}_{\widehat{T}_{\theta_t}\Gamma} \nabla \mathcal{L}_{|u_{\theta_t}} \right),$$

does not require to estimate a Gram matrix. More precisely, we have:

$$d\big((D,B) \circ u\big)^{\dagger}_{\theta_t} \left(\Pi^{\perp}_{\widehat{T}_{\theta_t}\Gamma} \nabla \mathcal{L}_{|u_{\theta_t}} \right) = \widehat{\phi}^{\dagger}_{\theta_t} \widehat{\nabla \mathcal{L}}_{\theta_t},$$

where: for all $1 \leqslant p \leqslant P, 1 \leqslant i \leqslant S$

- $\widehat{\phi}_{\theta_t\, i,p} := \left(\partial_p D\left[u_{\theta_t}\right](x_{i1}), \partial_p B\left[u_{\theta_t}\right](x_{i2})\right)$
- $\widehat{\nabla \mathcal{L}}_{\theta_t\, i} := \nabla \mathcal{L}_{|u_{\theta_t}}(x_i)$