Kernelization of Natural Gradient Methods for Physics Informed Neural Networks

MILES Seminar

Nilo Schwencke

TAU Team−INRIA Saclay; A&O−LISN−Paris-Saclay University; CNRS

November 3, 2025

Physics informed neural networks (PINNs)

Problem statement

We aim to solve:

$$\begin{cases} D(u) = f \in \mathrm{L}^2(\Omega \to \mathbb{R}, \mu) & \text{in } \Omega \\ B(u) = g \in \mathrm{L}^2(\partial\Omega \to \mathbb{R}, \sigma) & \text{on } \partial\Omega \end{cases}.$$

PINNs key idea

Optimize a neural network $u_{|\boldsymbol{\theta}}$ on the loss:

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

using autodiff to compute D, B (Raissi et al., 2019).

Problem

This leads to low accuracy with Adam.

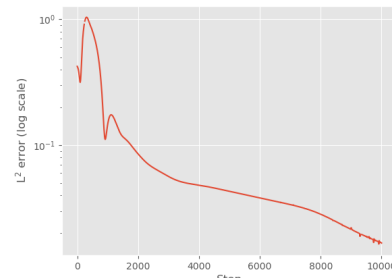


Figure: PINN solution under standard Adam optimization, to Laplace equation in 2 D.

Why does it work so bad ?

Intuition from NTK

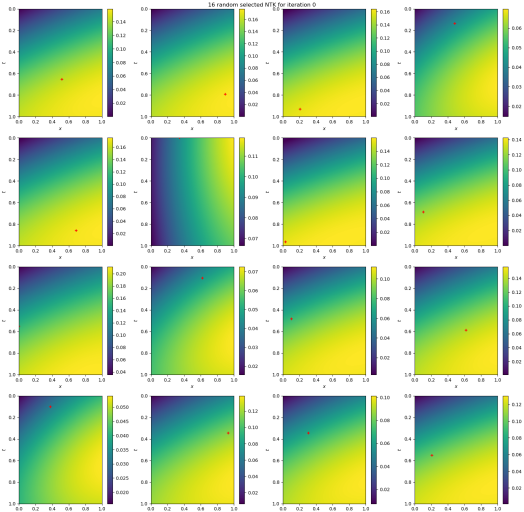


Figure: NTK for Laplace equation.

Intuition from Fourier

Consider the Fourier partial series:

$$S_N : \begin{cases} \mathbb{C}^{[\![-N,N]\!]} & \to & \mathrm{L}^2([0,1]) \\ (\alpha_k) & \mapsto & \sum_{k=-N}^{N} \alpha_k e^{2i\pi kx} \end{cases} .$$

S_N singular values are all 1 (perfect conditionning of the spectrum). **BUT:**

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

Conclusion: Differential operators strongly impact the spectral conditionning.

Natural Gradient

Reinterpreting quadratic loss

Consider the loss of a classical quadratic regression problem, with batch (x_i):

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

In the population limit:

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

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

A functional geometry perspective

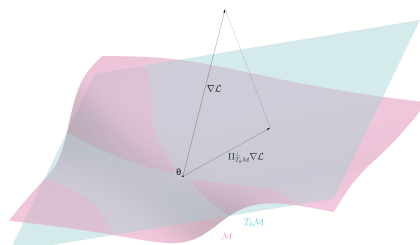
Natural gradient in functional space

The functional space is constrained to:

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

The Natural Gradient is then defined as:

$$\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t - \eta\, \mathrm{d}u^{\dagger}_{|\boldsymbol{\theta}_t}\left(\Pi^{\perp}_{T_{\boldsymbol{\theta}_t}\mathcal{M}}\nabla\mathcal{L}_{|u_{|\boldsymbol{\theta}_t}}\right),$$



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.

Definition-Proposition (Schwencke and Furtlehner (2025))

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

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

Corollary

The Natural Gradient update rewrites: $\boldsymbol{\theta}_{t+1} \leftarrow \boldsymbol{\theta}_t - \eta \, G^{\dagger}_{\boldsymbol{\theta}_t} \nabla \ell(\boldsymbol{\theta}_t)$; $\ell(\boldsymbol{\theta}) := \mathcal{L}(u_{|\boldsymbol{\theta}})$.

Shortcomings

- Computation of the Gram matrix $G_{\boldsymbol{\theta}_t}$ is quadratic in the number of parameters.
- Inversion of $G_{\boldsymbol{\theta}_t}$ is cubic

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

empirical Natural Gradient

(N)NTK in a nutshell

The functionnal dynamic of (N)GD on the empirical loss $\widehat{\ell}$ is described by (Jacot et al., 2018):

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

Key Idea

The empirical dynamics takes place in:

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

We can define the empirical Natural Gradient:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta\, \mathrm{d}u^{\dagger}_{|\boldsymbol{\theta}_t}\left(\Pi^{\perp}_{\widehat{T}_{\boldsymbol{\theta}_t}\mathcal{M}} \nabla\mathcal{L}_{|u_{|\boldsymbol{\theta}_t}}\right).$$

Byproduct

Yields an optimal criterion for (x_i) choice:

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

empirical Natural Gradient

Theorem (ANaGRAM)

Under mild assumptions:

$$du^{\dagger}_{|\boldsymbol{\theta}_t}\left(\Pi^{\perp}_{\widehat{T}_{\boldsymbol{\theta}_t}\mathcal{M}} \nabla\mathcal{L}_{|u_{|\boldsymbol{\theta}_t}}\right) = \widehat{\phi}^{\dagger}_{\boldsymbol{\theta}_t} \widehat{\nabla\mathcal{L}}_{\boldsymbol{\theta}_t},$$

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

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

Key fact

$\hat{\phi}^{\dagger}_{\boldsymbol{\theta}_t}$ **can be computed with a SVD, with complexity** $\mathrm{O}\big(\min(PS^2, P^2S)\big)$.

Corollary

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

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

Application to PINNs

Key remark

The only difference between the losses:

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

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

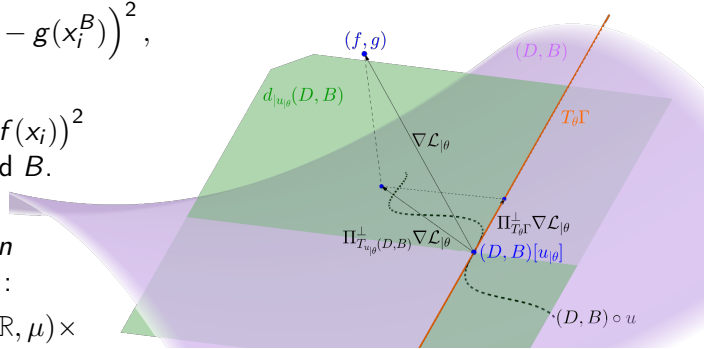
Proposition

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

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

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

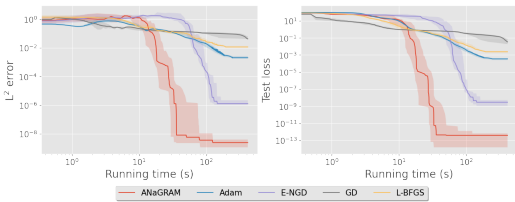


Figure: Performance comparison w.r.t running time for Laplace equation in 2 D:

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

1+1 D Heat equation

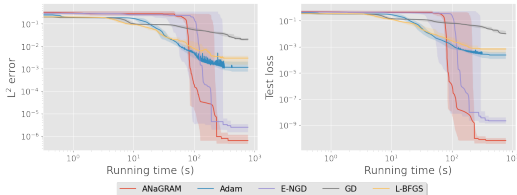


Figure: Performance comparison w.r.t running time for Heat equation in 1+1 D:

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

5 D Laplace equation

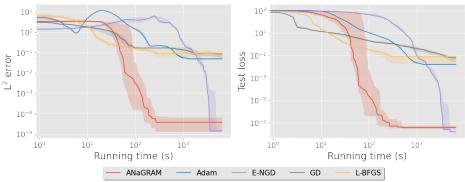


Figure: Performance comparison w.r.t running time for Laplace equation in 5 D:

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

1+1 D Allen-Cahn equation

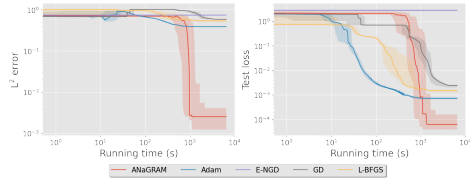
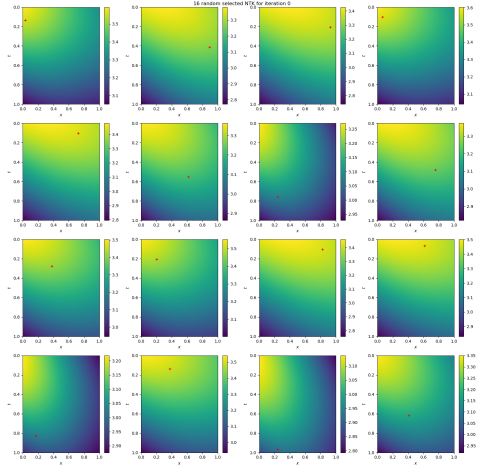
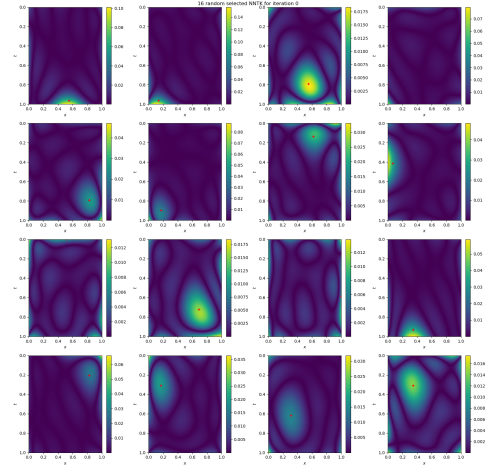


Figure: Performance comparison w.r.t running time for Allen-Cahn equation in 1+1 D:

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

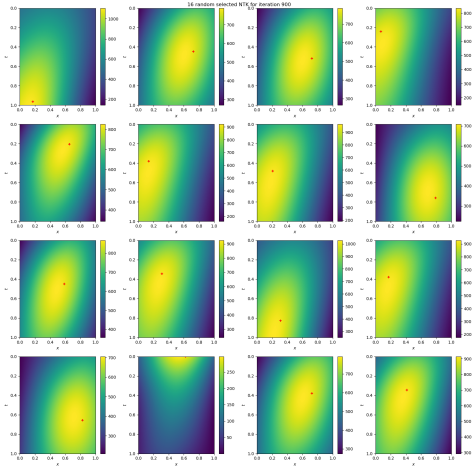


(a) NTK

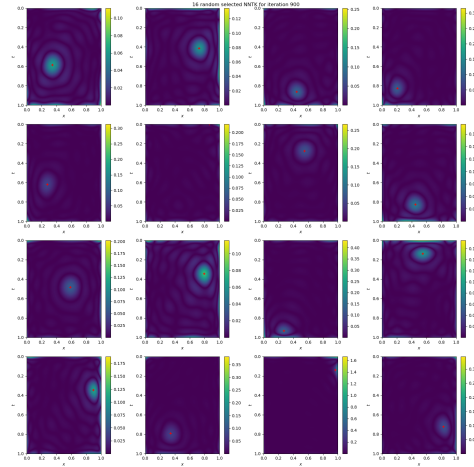


(b) NNTK

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

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

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:

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

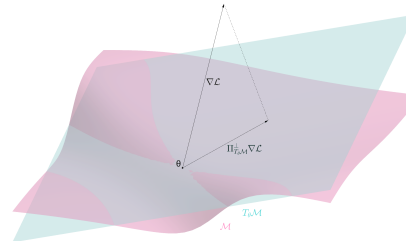
with $r_\alpha := \#\{i : \widehat{\Delta}_i \geqslant \alpha\} \leqslant \min(P, S)$.

Reconstruction Error (RCE)

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

Intuition on RCE

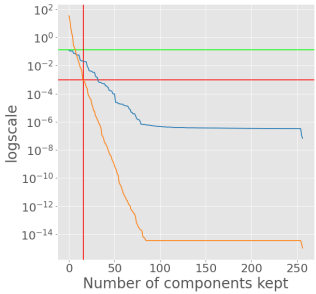
RCE_n accounts for the part of $\nabla\mathcal{L}_{\boldsymbol{\theta}}$ orthogonal to the n most important components of $\widehat{T}_{\boldsymbol{\theta}}\mathcal{M}$.



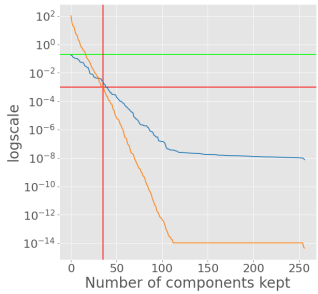
Remark

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

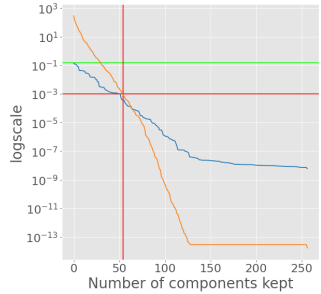
Empirical insights on the *cutoff* impact in *ANaGRAM*



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

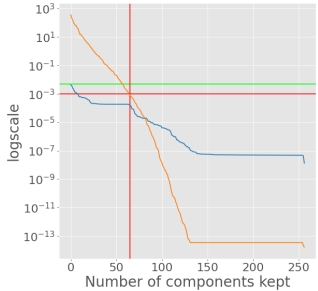


(b) Iteration 40: intersection point shifts rightward toward cut-off.

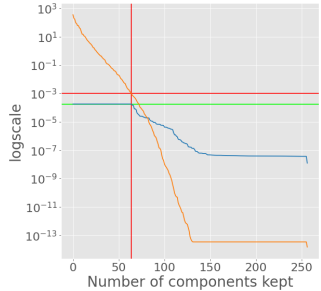


(c) Iteration 90: intersection point passes the cutoff threshold.

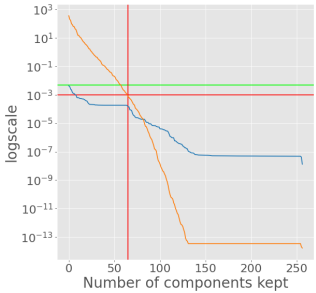
Empirical insights on the *cutoff* impact in *ANaGRAM*



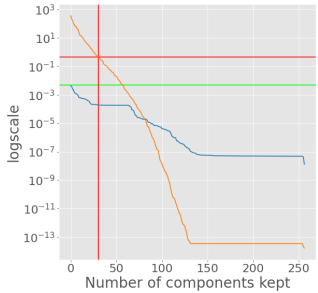
(d) Iteration 120. Beginning of *flattening*: RCE stabilizes at constant level before cutoff.



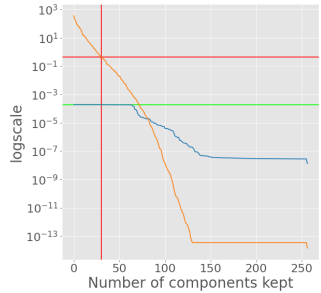
(e) Iteration 150: End of the *flattening phenomenon*. The train loss reaches the flattened part of the RCE.



(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: $\boldsymbol{\theta}_0 \in \mathbb{R}^P$

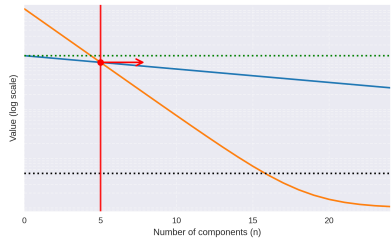
Precision target: $\epsilon > 0$

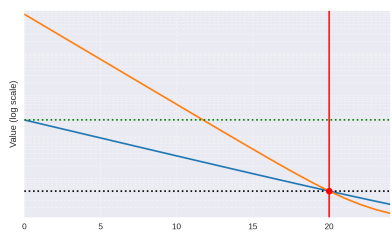
repeat

- $\widehat{U}_t, \widehat{\Delta}_t, \widehat{V}_t \leftarrow \mathrm{SVD}(\widehat{\phi}_t)$
- 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: $\boldsymbol{\theta}_t$





Algorithm 1: *AMStramGRAM* (sketch)

Input:

Initial parameters: $\boldsymbol{\theta}_0 \in \mathbb{R}^P$

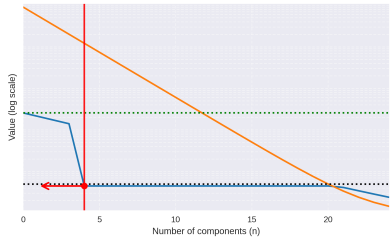
Precision target: $\epsilon > 0$

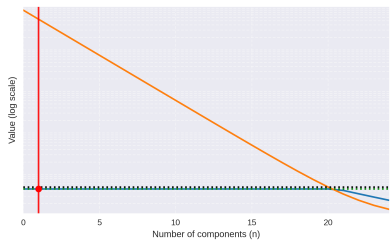
repeat

- $\widehat{U}_t, \widehat{\Delta}_t, \widehat{V}_t \leftarrow \mathrm{SVD}(\widehat{\phi}_t)$
- 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: $\boldsymbol{\theta}_t$





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

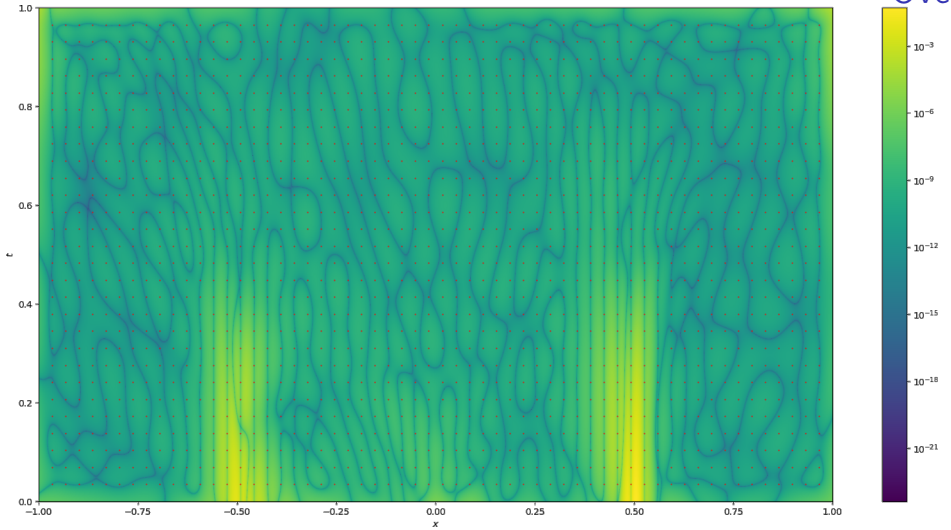


Figure: Overfitting on Allen–Cahn: residual lines align with sampling lines.

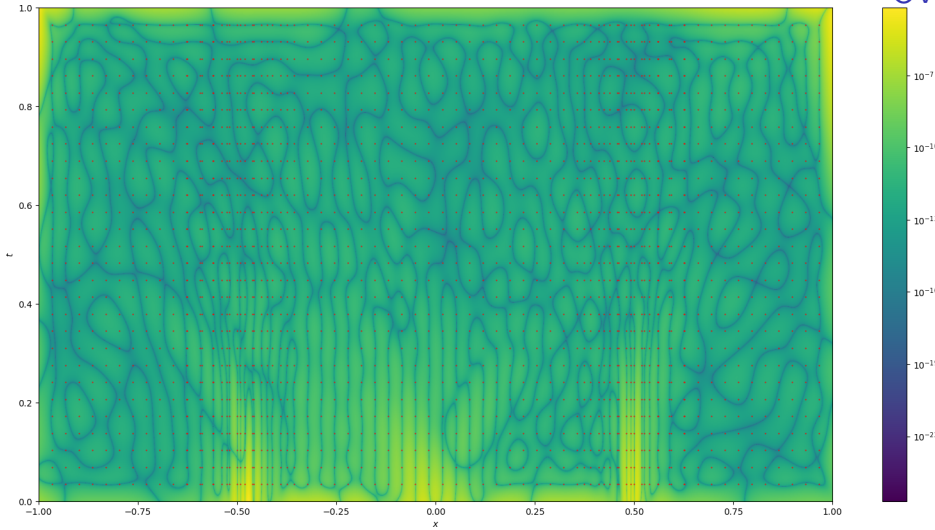


Figure: Overfitting on Allen–Cahn: densifying the sampling in overfitted regions mitigates overfitting.

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 $\mathcal{H}_0 \subset \mathcal{H}$ 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 $x, y \in \Omega$

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

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 $\boldsymbol{\theta} \in \mathbb{R}^P$, the generalized Green's function of D on $T_{\boldsymbol{\theta}}\mathcal{M} = \operatorname{Im} du_{|\boldsymbol{\theta}}$ is given by: for all $x, y \in \Omega$

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

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

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

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

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

Conclusion and Perspectives

Conclusions

- Anagram gives a theoretically founded simplication to any natural-gradient algorithm lowering the complexity from $O(P^3)$ to $O(\min(PN^2, P^2N))$, which is above stochastic gradient descent only by a factor $\min(P, N)$.
- 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.
- Establish theoretical connections with classical algorithms, such as FEMs, FDMs, *etc.*
- 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 !

Jacot, A., Gabriel, F., and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31.

Raissi, M., Perdikaris, P., and Karniadakis, G. (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.

Schwencke, N. and Furtlehner, C. (2025). ANaGRAM: A natural gradient relative to adapted model for efficient PINNs learning.

Schwencke, N., Rousselot, C., Shilova, A., and Furtlehner, C. (2025). AMStramGRAM: : Adaptive multi-cutoff strategy modification for anagram.

Urbán, J. F., Stefanou, P., and Pons, J. A. (2025). Unveiling the optimization process of physics informed neural networks: How accurate and competitive can PINNs be? *Journal of Computational Physics*, 523:113656.

Illustration of natural gradient dynamics

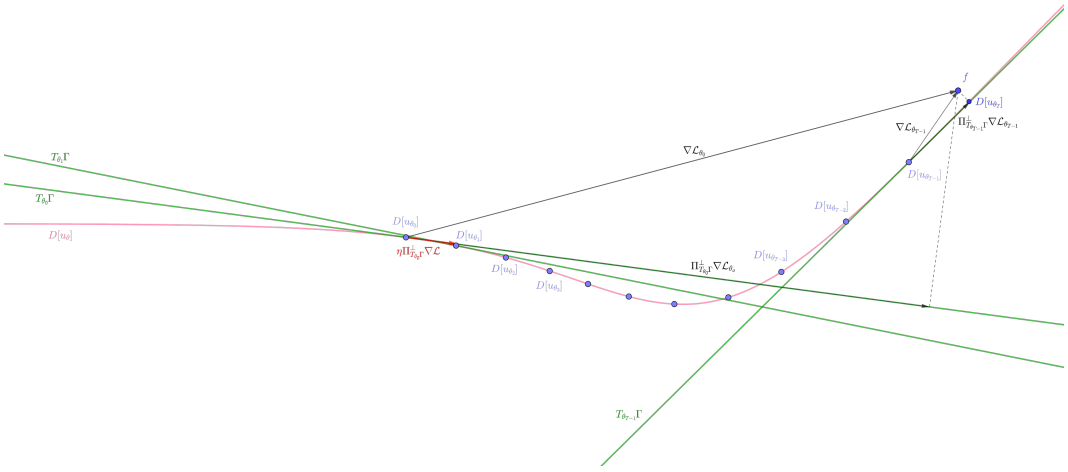


Figure: Illustration of PINNs learning process under natural gradient, as successive applications of Green's function

Corollary

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

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

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

Theorem (ANaGRAM for PINNs)

Under mild assumptions, the empirical natural gradient update:

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

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

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

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

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