Coupling-Robust Accuracy in Multiphysics PINNs via Kronecker-Preconditioned Optimization

Youngjae Park [1] **Jaemin Kim** [2] **Junghwa Hong** [1]

Abstract

Physics-informed neural networks (PINNs) for coupled multiphysics systems suffer systematic accuracy degradation as inter-equation coupling strengthens. We systematically evaluate how PINN accuracy scales with coupling strength and show that combining the Kronecker-preconditioned optimizer SOAP with inverse-gradient-norm loss balancing (SOAP+GN) yields coupling-robust accuracy. Across 234 experiments spanning three 1D systems of increasing nonlinearity and a 2D electroosmotic flow benchmark, SOAP+GN maintains final-epoch L_2 degradation $\leq 1.1\times$ (ratio of strong- to weak-coupling error) even as coupling parameters vary over one to two orders of magnitude, compared with $> 10^2\times$ for Adam+GN. In the nonlinear Nernst–Planck–Poisson system, L_∞-norm balancing (LRA) catastrophically fails due to weight explosion to $O(10^9)$, whereas GN remains stable—revealing a structural limitation of LRA for nonlinearly coupled PDEs. SOAP+GN further scales to a 2D, 6-PDE electroosmotic flow system at EDL-resolved conditions—a regime that all prior PINN electrokinetics studies have avoided through simplified physics—where Adam+GN fails entirely ($L_2 > 0.9$).

1. Introduction

Physics-informed neural networks (PINNs) (Raissi et al., 2019) embed PDE residuals into neural network loss functions, enabling mesh-free solutions. While PINNs have achieved broad success on single-equation problems (Karniadakis et al., 2021; Wang et al., 2022; Hao et al., 2024), coupled multiphysics systems—where multiple PDEs interact through shared variables—remain challenging (Sun et al., 2024).

A fundamental difficulty is that coupled PDE residuals compete for optimizer capacity. Each loss term's gradient may conflict with others, and relative magnitudes depend on coupling strength. As coupling intensifies, this imbalance worsens, causing accuracy degradation far beyond the single-equation setting. Despite extensive work on adaptive weighting—both global loss balancing (Wang et al., 2021) and local pointwise schemes (McClenny and Braga-Neto, 2023; Anagnostopoulos et al., 2024)—and characterization of PINN failure modes in single-equation settings (Krishnapriyan et al., 2021), no prior work has systematically measured how accuracy scales with coupling strength across multiple systems.

Contributions. (1) We systematically evaluate how PINN accuracy scales with coupling strength across multiple systems, using the ratio of strong- to weak-coupling error as a diagnostic. (2) Through a 6-method factorial comparison (2 optimizers $\times$ 3 balancing schemes), we show that Kronecker preconditioning (SOAP) is a necessary condition for coupling robustness, while gradient balancing becomes increasingly essential as coupling nonlinearity grows. (3) We identify a structural failure mode of L_∞-norm balancing (LRA) in nonlinearly coupled systems, where adaptive weights diverge to $O(10^9)$ and suppress boundary conditions entirely. (4) We demonstrate SOAP+GN on a 2D, 6-PDE electroosmotic flow system that directly resolves the electric double layer (EDL)—a regime all prior PINN electrokinetics studies have avoided through simplified physics (Boltzmann equilibrium, slip BCs, or omission of flow coupling).

2. Background

2.1. Multiphysics PINNs

Consider N coupled PDEs $\mathcal{R}_i[\hat{u}_1, \ldots, \hat{u}_N](x) = 0$, $i = 1, \ldots, N$, with boundary conditions $\mathcal{B}_j[\hat{u}_1, \ldots, \hat{u}_N](x) = 0$ on $\partial\Omega$. Each field variable is approximated by an independent network $\hat{u}_i(x; \theta_i)$ with disjoint parameters ($\theta_i \cap \theta_j = \emptyset$ for $i \neq j$), following Wang et al. (2022) and Sun et al.

[1]Dept. of Control and Instrumentation Engineering, Korea University, Sejong, South Korea [2]BK21 FOUR Smart Mobility Education and Research Team, Korea University, Sejong, South Korea. Correspondence to: Junghwa Hong <hongjh32@korea.ac.kr>.

Proceedings of the AI4Physics Workshop at the 43rd International Conference on Machine Learning (AI4Physics@ICML 2026), Seoul, South Korea. 2026. Copyright 2026 by the author(s).

(2024). The training loss is

$$\mathcal{L}(\theta) = \sum_i \lambda_i \frac{1}{N_r} \sum_{k=1}^{N_r} \mathcal{R}_i(x_k^r)^2 + \sum_j \lambda_j^{\mathrm{b}} \frac{1}{N_b} \sum_{k=1}^{N_b} \mathcal{B}_j(x_k^b)^2, \tag{1}$$

where $\{x_k^r\}$ and $\{x_k^b\}$ are interior collocation and boundary points, respectively, and λ_i, λ_j^{b} are loss weights determined by the balancing scheme (fixed at 1 when no balancing is applied; see Section 2.3).

2.2. SOAP Optimizer

SOAP (Vyas et al., 2024) applies Kronecker-factored preconditioning: $P \approx (L \otimes R)^{-1/2}$, where L and R are running gradient outer-product averages per weight matrix. Wang et al. (2025) showed SOAP achieves state-of-the-art single-PDE PINN accuracy across 10 benchmarks, but all benchmarks involve single-physics systems (reaction–diffusion, Navier–Stokes); its effectiveness on cross-domain multiphysics coupling remains unverified. Intuitively, the Kronecker preconditioner adapts to each weight matrix's effective curvature, providing natural normalization when coupling causes PDE residual gradients to differ by orders of magnitude. This gap motivates the present study.

2.3. Gradient Balancing

When no balancing is applied, all loss weights are fixed at $\lambda_i = 1$.

Inverse gradient norm (GN). Following Wang et al. (2025, Appendix G), we set $\lambda_i \propto 1/\|\nabla_\theta \mathcal{L}_i\|_2$ with exponential moving average (momentum 0.9, update frequency 1000 steps). This enforces $\|\lambda_i \nabla \mathcal{L}_i\|_2 = \bar{m}$ for all i. We adopt GN as the primary balancing scheme because this equal gradient-norm condition is required for the single-step convergence bound of preconditioned gradient descent to be tight (Wang et al., 2025).

Learning Rate Annealing (LRA). Wang et al. (2021) proposed L_∞-norm balancing: $\lambda_i \propto 1/\max_j |\partial \mathcal{L}_i / \partial \theta_j|$, updated every step with EMA ($\alpha = 0.1$). This equalizes maximum gradient components but not L_2 norms.

3. Benchmark Systems

3.1. 1D Thermoelasticity (Linear, One-Way Coupling)

$$-\kappa T'' = f_T, \qquad -E\,u'' + \gamma\,T' = f_u \tag{2}$$

on $x \in [0,1]$ with Dirichlet BCs. The coupling is one-way: temperature T drives displacement u through the thermal stress term $\gamma T'$, but u does not affect T. Manufactured exact solutions $T^*(x) = \sin(\pi x) + 1$, $u^*(x) = \sin(2\pi x)$ determine the source terms $f_T = \kappa \pi^2 \sin(\pi x)$, $f_u = E(2\pi)^2 \sin(2\pi x) + \gamma\pi\cos(\pi x)$ analytically. Coupling parameter $\gamma \in \{1, 5, 10, 25, 50\}$; $\kappa = E = 1$.

3.2. 1D Reaction–Diffusion (Linear, Bidirectional)

$$\begin{aligned} D_1\, c_A'' - k_1 c_A + k_2 c_B &= f_1 \\ D_2\, c_B'' + k_1 c_A - (k_2 + k_3) c_B + k_4 c_C &= f_2 \\ D_3\, c_C'' + k_3 c_B - k_4 c_C &= f_3 \end{aligned} \tag{3}$$

on $x \in [0,1]$ with Dirichlet BCs. The reversible chain $A \rightleftharpoons B \rightleftharpoons C$ creates bidirectional coupling: each species' concentration appears in the reaction terms of its neighbors. We set symmetric rates $k_1 = k_2 = k_3 = k_4 = k$ with $k \in \{1, 5, 10, 30\}$; $D_i = 1$ for all species. Manufactured exact solutions $c_A^* = \sin(\pi x) + 2$, $c_B^* = \cos(\pi x) + 2$, $c_C^* = \sin(2\pi x) + 2$ determine source terms f_i.

3.3. 1D Nernst–Planck–Poisson (Nonlinear)

$$\begin{aligned} c_+'' + (c_+ \phi')' &= 0 \\ c_-'' - (c_- \phi')' &= 0 \\ \varepsilon^2 \phi'' + (c_+ - c_-) &= 0 \end{aligned} \tag{4}$$

on $x \in [0,1]$ with Boltzmann equilibrium BCs: $\phi(0) = \zeta$, $c_\pm(0) = e^{\mp\zeta}$; $\phi(1) = 0$, $c_\pm(1) = 1$; where $\zeta = 1.0$. The exact solution is the 1D Poisson–Boltzmann equilibrium, obtained by solving the nonlinear BVP $\varepsilon^2 \phi'' = 2\sinh(\phi)$ numerically (`solve_bvp`), with $c_\pm^* = e^{\mp\phi^*}$. Coupling is nonlinear and circular: ion concentrations $c_\pm$ enter Poisson through charge density, while ϕ enters NP through electromigration fluxes $c_\pm \phi'$. The parameter ε (= Debye length / domain size) $\in \{1.0, 0.5, 0.2\}$ controls coupling stiffness; as $\varepsilon \to 0$ the Poisson equation becomes singularly perturbed, creating sharp boundary layers in ϕ and $c_\pm$.

3.4. 2D Electroosmotic Flow (6 PDEs, 4 Networks)

Nernst–Planck + Poisson + Stokes system:

$$\begin{aligned} \varepsilon^2 \nabla^2 \phi + (c_+ - c_-) &= 0 \\ \nabla^2 c_\pm \pm \nabla \cdot (c_\pm \nabla \phi) &= 0 \\ -\nabla p + \mu \nabla^2 \mathbf{u} + E_x (c_+ - c_-)\hat{x} &= 0 \\ \nabla \cdot \mathbf{u} &= 0 \end{aligned} \tag{5}$$

Four independent networks: ϕ-net, c^+-net, c^--net, flow-net(u, v, p). EDL-resolved (no Helmholtz–Smoluchowski slip), $\zeta = 2.0$, $\varepsilon \in \{0.2, 0.1, 0.05\}$, $E_x = \mu = 1$. Boundary conditions: walls ($y = 0, 1$): $\phi = \zeta$, $c_\pm = e^{\mp\zeta}$, $u = v = 0$ (no-slip); channel ends ($x = 0, 1$): Dirichlet from reference solution, $p = 0$. The reference solution is the fully developed 1D Poisson–Boltzmann equilibrium (y-dependent BVP solved via `solve_bvp`), used as

an x-independent exact solution on the 2D domain. Only SOAP+GN and Adam+GN evaluated (top two from 1D).

EDL-resolved simulation of electrokinetic flows poses well-known numerical challenges due to the singular perturbation structure of the Poisson equation (Mani & Wang, 2020). This benchmark is significantly more challenging than prior PINN electrokinetics studies, all of which employ one or more simplifications to avoid the ε^2-stiffness inherent in EDL-resolved simulations (Table 1).

Our formulation solves the full NP+Poisson+Stokes system with bidirectional nonlinear coupling ($c_\pm \nabla\phi$ in NP, $(c_+ - c_-)$ in both Poisson and Stokes) at $\varepsilon = 0.05$ ($\varepsilon^2 = 0.0025$), where the Debye layer occupies 5% of the channel width.

4. Experimental Setup

Architecture. Each field variable is approximated by an independent MLP with disjoint parameters, following the segregated-network design of Sun et al. (2024) (cf. Section 2), where $\hat{\cdot}$ denotes the network approximation of the corresponding exact field variable. Figure 1 illustrates the general training pipeline and summarizes the per-system network configurations. The network approximations for each system are:

Thermoelasticity (2 networks): $\hat{T}(x;\theta_T)$, $\hat{u}(x;\theta_u)$. *Reaction–Diffusion* (3 networks): $\hat{c}_A(x;\theta_A)$, $\hat{c}_B(x;\theta_B)$, $\hat{c}_C(x;\theta_C)$. *NP+P* (3 networks): $\hat{c}_+(x;\theta_+)$, $\hat{c}_-(x;\theta_-)$, $\hat{\phi}(x;\theta_\phi)$. *2D NPPS* (4 networks): $\hat{\phi}(\mathbf{x};\theta_\phi)$, $\hat{c}_+(\mathbf{x};\theta_+)$, $\hat{c}_-(\mathbf{x};\theta_-)$, $[\hat{u},\hat{v},\hat{p}](\mathbf{x};\theta_f)$; where $\mathbf{x} = (x, y)$ and all parameter sets are pairwise disjoint: $\theta_i \cap \theta_j = \emptyset$ for $i \neq j$. In the 2D system, the Stokes variables (u, v, p) share a single network (flow-net) with output dimension 3; all other networks have scalar output. All networks use tanh activation and Xavier uniform initialization.

Collocation. Uniform sampling for Thermo and RD (200 interior points). For NP+P and 2D NPPS, half uniform and half concentrated near the boundary layer to resolve the Debye layer (cosine distribution for 1D; uniform random for 2D). Interior points: 300 for 1D; 3000 for 2D. Boundary points for 2D: 200 per edge (800 total).

Methods. Six configurations for 1D; two (SOAP+GN, Adam+GN) for 2D. Learning rate (η): 10^{-3} (Thermo, RD); 3×10^{-4} (NP+P, 2D NPPS).

Training. 30,000 epochs (1D), 50,000 epochs (2D Adam+GN), 100,000 epochs (2D SOAP+GN; Adam+GN failed to reduce L_2 below 0.9 throughout 50,000 epochs, with persistent oscillation indicating inability to escape the spurious quasi-neutral solution described in Section 5.3; see Appendix A.5). Three seeds per configuration. Total: 234 runs (90 Thermo + 72 RD + 54 NP+P + 18 2D). SOAP momentum parameters: $\beta = (0.99, 0.999)$ for Thermo and RD, $\beta = (0.95, 0.95)$ for NP+P and 2D NPPS; preconditioner update frequency 2.

Metric. Final-epoch relative L_2 error, averaged across field variables and seeds.

Coupling degradation ratio. To quantify coupling robustness, we define $D = \bar{L}_2(\text{strong})/\bar{L}_2(\text{weak})$, where $\bar{L}_2$ is the 3-seed mean. Values near 1 indicate coupling robustness. The ratio is meaningful only for methods that converge (Final $L_2 < 0.1$) with sufficient reproducibility; we mark non-converged methods as FAIL and flag high-variance ratios.

5. Results

5.1. Main Result: Coupling Robustness

Table 3 reports coupling degradation ratios under the final-epoch metric.

Figure 2 visualizes these trends: SOAP+GN (green) maintains a flat profile across all coupling strengths, while Adam-based methods degrade or fail entirely in NP+P. SOAP+GN is the only method achieving degradation $\leq 1.1\times$ across all three systems. Three methods fail entirely in NP+P (Final $L_2 > 0.1$), leaving only Adam, SOAP, and SOAP+GN with computable degradation ratios in the nonlinear regime. The degradation ratio alone, however, does not capture absolute accuracy; Table 4 reports Final L_2 at extreme coupling.

SOAP+GN achieves the most consistent accuracy across all regimes. In linear systems, SOAP alone and SOAP+LRA achieve comparable or occasionally lower absolute L_2 (e.g., SOAP+LRA Thermo $\gamma = 25$: 3.9e-5 vs SOAP+GN: 1.0e-4), but both degrade in NP+P. Adam-based methods with balancing (Adam+LRA, Adam+GN) maintain L_2 in the 10^{-4}–10^{-3} range for linear systems but fail catastrophically in NP+P ($L_2 > 0.3$ at strong coupling), confirming that balancing without preconditioning cannot rescue nonlinear coupling.

5.2. Factor Decomposition

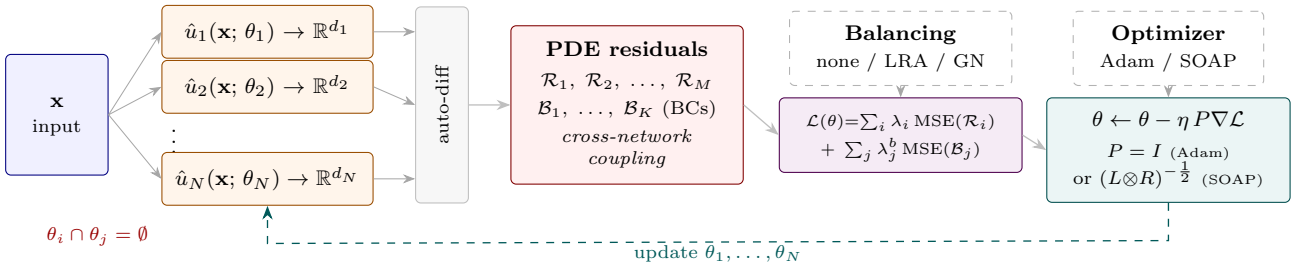
The 6-method factorial design decomposes coupling robustness into two factors:

Preconditioning (SOAP). Comparing Adam vs SOAP (both unbalanced) at the strongest coupling: SOAP improves Final L_2 by $143\times$ in Thermo ($\gamma = 50$: 1.4e-2 $\rightarrow$ 9.8e-5), $138\times$ in RD ($k = 30$: 1.1e-2 $\rightarrow$ 8.0e-5), and $41\times$ in NP+P ($\varepsilon = 0.2$: 3.8e-2 $\rightarrow$ 9.3e-4). SOAP is a necessary condition—no Adam-based method achieves consistent robustness.

Gradient balancing (GN). The relative importance of GN increases with coupling nonlinearity. *Linear systems*: SOAP alone yields Final degradation $1.2\times$ (Thermo), $0.8\times$ (RD)—

Table 1. Prior neural network studies on electrokinetic flows.

Study	Gov. eqs.	Flow coupling	EDL treatment	ε regime
Sun et al. (2024)	NP+P+NS (HS slip)	Bidir. (bulk); HS slip	Avoided via slip BC	10^{-2}; bypassed
Tao et al. (2025)	NS only (const. ρ_e)	N/A	Not modeled	N/A
Huang et al. (2025)	NP + Poisson	None	Debye-scaled ($\varepsilon^2=1$)	1.0
Merdasi et al. (2023)	PB+Lap+NS+energy	One-way (PB→NS)	Boltzmann equil.	Thin EDL
Cai et al. (2021)	NP+P+Stokes	Bidir. (full)	Resolved by conv. solver (data-driven)	10^{-2}
Ours (PINN)	**NP+P+Stokes**	**Bidir. (full)**	**Directly resolved (data-free)**	**0.05–0.2**



System	N	Networks (output dim)	Depth × Width	Input
Thermoelasticity	2	$\hat{T}$, $\hat{u}$ (each $\mathbb{R}^1$)	4 × 64	$x \in \mathbb{R}^1$
Reaction–Diffusion	3	$\hat{c}_A$, $\hat{c}_B$, $\hat{c}_C$ (each $\mathbb{R}^1$)	4 × 64	$x \in \mathbb{R}^1$
NP+P	3	$\hat{c}_+$, $\hat{c}_-$, $\hat{\phi}$ (each $\mathbb{R}^1$)	4 × 64	$x \in \mathbb{R}^1$
2D NPPS	4	$\hat{\phi}$, $\hat{c}_+$, $\hat{c}_-$ ($\mathbb{R}^1$); $[\hat{u}, \hat{v}, \hat{p}]$ ($\mathbb{R}^3$)	5 × 128	$\mathbf{x} \in \mathbb{R}^2$

Figure 1. Segregated-network PINN architecture with configurable optimizer and loss balancing. Top: general N-network training pipeline. Each field variable is approximated by an independent MLP with disjoint parameters ($\theta_i \cap \theta_j = \emptyset$). Outputs are differentiated via automatic differentiation (AD) to form PDE residuals, which are combined into a weighted loss $\mathcal{L}$. The loss balancing scheme (none, LRA, or GN) determines the weights λ_i; the optimizer (Adam or SOAP) determines the preconditioner P. Bottom: per-system network configurations.

Table 2. Method configurations.

Method	Optimizer	Balancing
Adam	Adam	None
SOAP	SOAP	None
Adam+LRA	Adam	LRA ($\alpha=0.1$)
SOAP+LRA	SOAP	LRA ($\alpha=0.1$)
Adam+GN	Adam	GN (freq=1000, momentum=0.9)
SOAP+GN	SOAP	GN (freq=1000, momentum=0.9)

already robust. GN adds marginal improvement (→ 1.1×, 0.9×). *Nonlinear NP+P*: SOAP alone degrades 10.9× (Final). Adding GN reduces this to 1.0×—GN is essential.

Quantitative hierarchy. At NP+P $\varepsilon = 0.2$ (Final L_2), switching optimizer (Adam → SOAP, no balancing) improves accuracy by 41× ($3.8 \times 10^{-2} \to 9.3 \times 10^{-4}$), while adding balancing to the same optimizer (SOAP → SOAP+GN) improves by 4× ($9.3 \times 10^{-4} \to 2.3 \times 10^{-4}$). Preconditioning contributes an order of magnitude more than balancing. Moreover, balancing is only effective on top of the right preconditioner: Adam+GN *worsens* to 3.7e-1 at $\varepsilon=0.2$, indicating that gradient norm equalization without curvature-adapted step sizes amplifies rather than resolves the optimization difficulty.

Table 3. Coupling degradation (Final L_2, 3-seed mean). Methods with Final $L_2 > 0.1$ at either endpoint are marked FAIL; degradation ratio is not computed.

Method	Thermo (γ:1→50)	RD (k:1→30)	NP+P (ε:1→0.2)
Adam	3.3×	3.2×	10.0×
SOAP	1.2×	**0.8**×	10.9×
Adam+LRA	0.2×†	18.7×	FAIL
SOAP+LRA	1.4×	2.6×	FAIL
Adam+GN	0.3×†	0.5×†	FAIL
SOAP+GN	**1.1**×	0.9×	**1.0**×

†Ratio < 1 with high seed variance at one or both endpoints (see Appendix A for per-configuration CV); the apparent improvement likely reflects noise rather than genuine coupling benefit.

5.3. LRA Failure in Nonlinear Coupling

Adam+LRA fails catastrophically in NP+P: Final $L_2 > 0.47$ at all ε, all seeds—up to ∼270× worse than unbalanced

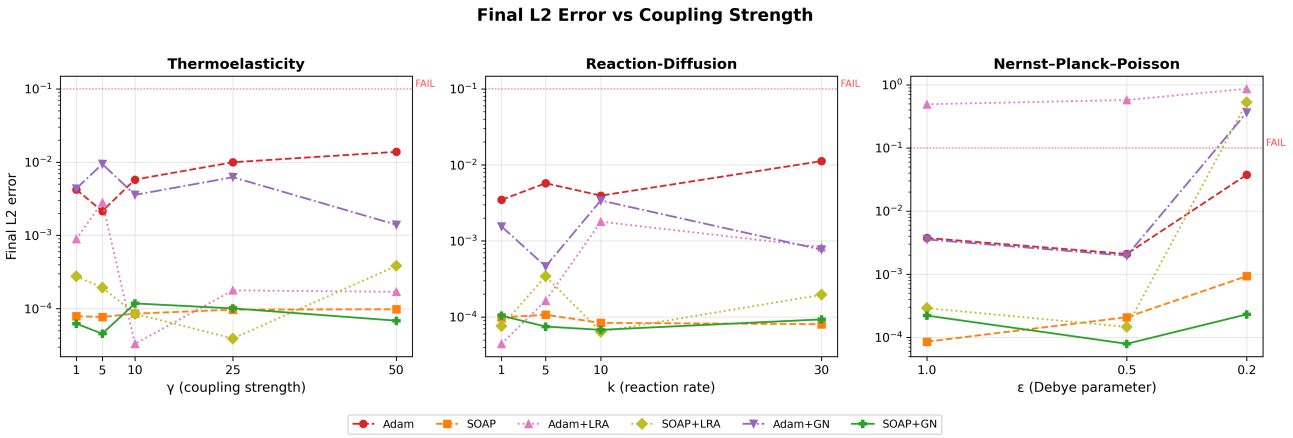


Figure 2. Final L_2 error vs. coupling strength across three benchmark systems. The dashed red line marks the FAIL threshold ($L_2 = 0.1$). SOAP+GN (green, solid) maintains flat accuracy across all coupling regimes, while Adam-based methods degrade by 1–3 orders of magnitude. In NP+P (right), Adam+LRA and SOAP+LRA exceed the FAIL threshold at all or most ε values.

Table 4. Absolute Final L_2 at weak and strong coupling (3-seed mean).

Method	Thermo		RD		NP+P	
	$\gamma=1$	$\gamma=50$	$k=1$	$k=30$	$\varepsilon=1$	$\varepsilon=0.2$
Adam	4.2e-3	1.4e-2	3.5e-3	1.1e-2	3.7e-3	3.8e-2
SOAP	7.9e-5	9.8e-5	9.9e-5	**8.0e-5**	**8.5e-5**	9.3e-4
Adam+LRA	8.9e-4	1.7e-4	**4.4e-5**	8.3e-4	4.9e-1	8.6e-1
SOAP+LRA	2.8e-4	3.8e-4	7.6e-5	2.0e-4	2.9e-4	5.3e-1
Adam+GN	4.3e-3	1.4e-3	1.6e-3	7.7e-4	3.6e-3	3.7e-1
SOAP+GN	**6.3e-5**	**6.9e-5**	1.0e-4	9.3e-5	2.2e-4	**2.3e-4**

Adam in individual seeds (e.g., mean $L_2 = 0.49$ vs 3.7×10^{-3} at $\varepsilon=1.0$; $\sim$130$\times$ in mean), indicating LRA actively harms optimization. The mechanism is structural: the NP+P equations are homogeneous in the unknowns, so any quasi-neutral constant state ($c_+ \approx c_- \approx \bar{c}$, $\phi' \approx 0$) satisfies PDE residuals to machine precision while violating BCs. Near this spurious solution, PDE gradient norms vanish and LRA weights grow to $O(10^9)$, trapping the optimizer. This failure is ε-independent ($L_2 > 0.46$ even at $\varepsilon=1.0$), confirming it is not stiffness-induced but a structural incompatibility between L_∞-norm balancing and homogeneous nonlinear PDEs. By contrast, the linear benchmarks have inhomogeneous source terms that prevent such spurious solutions. GN mitigates this through two mechanisms: (1) the mean gradient norm $\bar{m}$ includes BC terms, preventing individual PDE weights from diverging independently, and (2) infrequent updates (every 1000 steps) with momentum smooth out transient gradient fluctuations near the spurious solution. SOAP+LRA delays but does not prevent the collapse (2/3 seeds fail at $\varepsilon=0.2$).

Table 5. 2D NPPS Final L_2 ($\zeta = 2.0$, 3-seed mean $\pm$ std).

ε	SOAP+GN	Adam+GN	Ratio
0.2	1.32e-3 $\pm$ 8.14e-4	9.12e-1 $\pm$ 7.41e-2	691$\times$
0.1	2.85e-3 $\pm$ 1.78e-3	1.07e+0 $\pm$ 9.76e-3	375$\times$
0.05	1.50e-2 $\pm$ 1.07e-2	1.27e+0 $\pm$ 4.36e-3	85$\times$

5.4. 2D Electroosmotic Flow

Table 5 shows Final L_2 on the 2D NP+P+Stokes system with $\zeta = 2.0$.

Adam+GN fails at all ε ($L_2 > 0.9$), while SOAP+GN achieves 10^{-3} accuracy at $\varepsilon = 0.2$. As noted in Table 1, no prior PINN electrokinetics work has attempted EDL-resolved NP+Poisson+Stokes with full bidirectional coupling; existing studies avoid the ε^2-stiffness through Boltzmann equilibrium assumptions (Merdasi et al., 2023), constant charge density simplification (Tao et al., 2025), slip boundary conditions (Sun et al., 2024), Debye-scaled coordinates that eliminate ε^2 (Huang et al., 2025), or omission of flow coupling. Data-driven neural operator approaches have also been explored for electroconvection (Cai et al., 2021), but require training data from conventional solvers. Our result at $\varepsilon=0.2$ ($L_2 = 1.3\times10^{-3}$) demonstrates that SOAP+GN makes this previously intractable regime accessible to PINNs.

At $\varepsilon=0.05$, SOAP+GN reaches $L_2 = 1.5\times10^{-2}$—an approximate rather than fully converged solution, suggesting room for improvement through hyperparameter tuning or extended training. Figure 3 illustrates the qualitative difference: SOAP+GN captures the EDL structure in ϕ with localized error in the channel interior, while Adam+GN produces a near-constant field with $O(1)$ error everywhere. Critically,

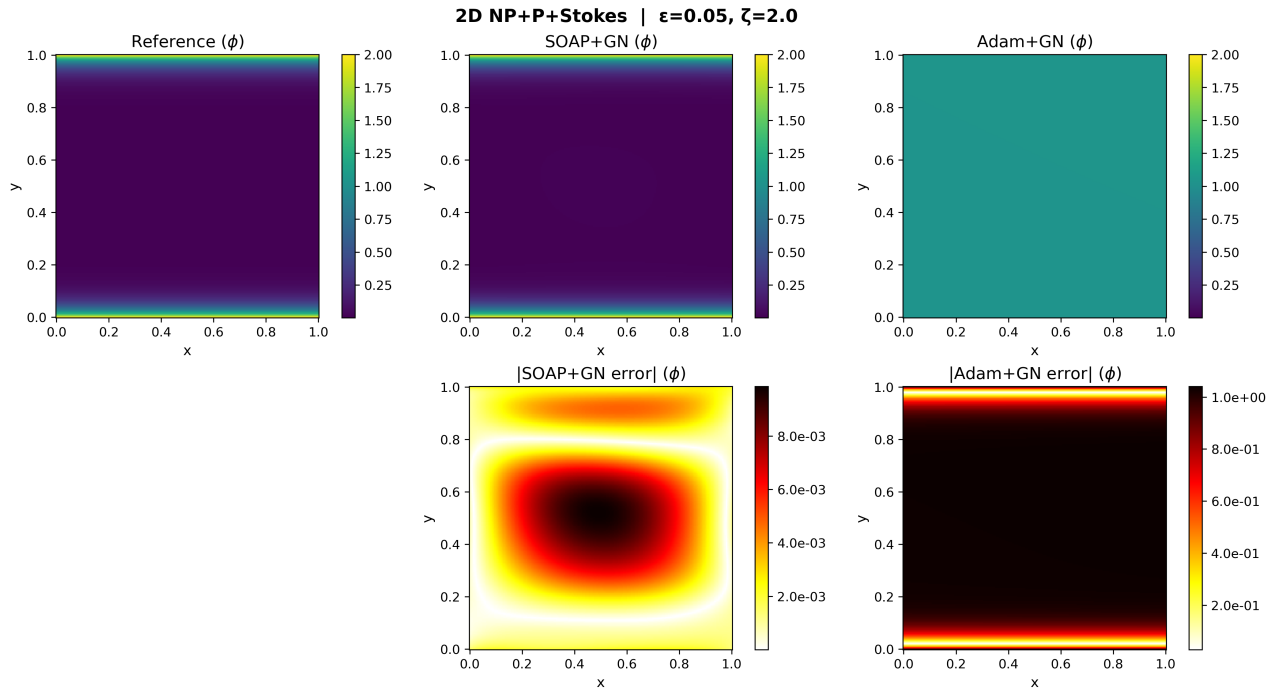


Figure 3. 2D NP+P+Stokes at $\varepsilon = 0.05$, $\zeta = 2.0$: electric potential ϕ field. Top row: reference solution (left), SOAP+GN prediction (center), Adam+GN prediction (right). Bottom row: pointwise absolute error for SOAP+GN ($\max \sim 10^{-2}$, left) and Adam+GN ($\max \sim 1.0$, right). SOAP+GN resolves the Debye layer at both walls; Adam+GN collapses to a near-constant field.

Adam+GN fails at all ε in this 2D system ($L_2 > 0.9$), confirming that Kronecker preconditioning is necessary to make EDL-resolved multiphysics PINNs tractable—the very regime that prior work avoided (Table 1).

6. Discussion and Future Work

Why SOAP enables coupling robustness. SOAP's Kronecker preconditioner $(L \otimes R)^{-1/2}$ adapts to the local curvature of each weight matrix. In coupled systems, the effective curvature changes as coupling strengthens—PDE residuals with ε^2-prefactored terms have suppressed gradient norms that standard Adam treats equally. SOAP's curvature-adapted step sizes can normalize these disparities without explicit loss weighting. Verifying this through block-wise NTK eigenvalue analysis is a key direction for future work; we provide an initial theoretical treatment in Appendix B.

Limitations. (1) Linear benchmarks use manufactured solutions. (2) NP+P is limited to $\varepsilon \geq 0.2$; singular perturbation regimes likely require Fourier features or adaptive sampling. (3) 2D results use only SOAP+GN vs Adam+GN; full 6-method comparison in 2D is left for future work. (4) GN hyperparameters (freq=1000, momentum=0.9) are adopted from Wang et al. (2025) without system-specific tuning. (5) In 2D, GN weights diverged to $O(10^{12})$ when one PDE converged much faster, suggesting weight clipping may be needed for larger systems.

Computational cost. SOAP incurs $\sim$1.6$\times$ wall-clock overhead per epoch compared to Adam (16 ms vs 10 ms backward pass on RTX 4060 Ti) due to Kronecker factor maintenance. Given the 40–100$\times$ accuracy improvement in coupled systems, this overhead is modest.

Future work. Achieving fully converged solutions at small ε through GN weight clipping, learning rate scheduling, and adaptive collocation is a natural next step. Extending coupling strength sweeps to time-dependent multiphysics systems is another direction.

7. Conclusion

We have demonstrated that SOAP+GN achieves coupling-robust accuracy in multiphysics PINNs: final-epoch degradation of 0.9–1.1$\times$ across three 1D systems with coupling variations of 5–50$\times$, compared with $> 10^2\times$ for Adam+GN. The preconditioning effect dominates by an order of magnitude over gradient balancing (41$\times$ vs 4$\times$ at NP+P $\varepsilon = 0.2$); however, preconditioning is necessary but not sufficient for nonlinear coupling, where GN provides an additional 4$\times$ improvement that closes the gap to coupling-invariant performance. This robustness enables SOAP+GN to solve a 2D, 6-PDE EDL-resolved electroosmotic flow problem—a regime that all prior PINN electrokinetics studies have avoided

through simplified physics—achieving $L_2 = 1.3 \times 10^{-3}$ at $\varepsilon = 0.2$ where Adam+GN fails entirely. We identify a failure mode of L_∞-norm balancing (LRA) in nonlinear systems and show that coupling strength sweeps provide a useful diagnostic for evaluating multiphysics PINN methods.

Code is available at https://github.com/YoungjaePark99/PINN_multiphysics.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Maziar Raissi, Paris Perdikaris, and George E Karniadakis. 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, 2019.

George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.

Sifan Wang, Xinling Yu, and Paris Perdikaris. When and why PINNs fail to train: A neural tangent kernel perspective. *Journal of Computational Physics*, 449:110768, 2022.

Zhongkai Hao, Jiachen Yao, Chang Su, Hang Su, Ziao Wang, Fanzhi Lu, Zeyu Xia, Yichi Zhang, Songming Liu, Lu Lu, et al. PINNacle: A comprehensive benchmark of physics-informed neural networks for solving PDEs. *Advances in Neural Information Processing Systems*, 37: 76721–76774, 2024.

Runze Sun, Hyogu Jeong, Jiachen Zhao, Yixing Gou, Emilie Sauret, Zirui Li, and Yuantong Gu. A physics-informed neural network framework for multi-physics coupling microfluidic problems. *Computers & Fluids*, 284:106421, 2024.

Sifan Wang, Yujun Teng, and Paris Perdikaris. Understanding and mitigating gradient flow pathologies in physics-informed neural networks. *SIAM Journal on Scientific Computing*, 43(5):A3055–A3081, 2021.

Levi D McClenny and Ulisses M Braga-Neto. Self-adaptive physics-informed neural networks. *Journal of Computational Physics*, 474:111722, 2023.

Sokratis J Anagnostopoulos, Juan Diego Toscano, Nikolaos Stergiopulos, and George Em Karniadakis. Residual-based attention in physics-informed neural networks. *Computer Methods in Applied Mechanics and Engineering*, 421:116805, 2024.

Aditi Krishnapriyan, Amir Gholami, Shandian Zhe, Robert Kirby, and Michael W Mahoney. Characterizing possible failure modes in physics-informed neural networks. *Advances in neural information processing systems*, 34: 26548–26560, 2021.

Nikhil Vyas, Depen Morwani, Rosie Zhao, Mujin Kwun, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. SOAP: Improving and stabilizing Shampoo using Adam. *arXiv preprint arXiv:2409.11321*, 2024.

Sifan Wang, Ananyae Kumar Bhartari, Bowen Li, and Paris Perdikaris. Gradient alignment in physics-informed neural networks: A second-order optimization perspective. *arXiv preprint arXiv:2502.00604*, 2025.

Ze Tao, Ke Xu, and Fujun Liu. LSTM-PINN: An hybrid method for prediction of steady-state electrohydrodynamic flow. *Journal of Computational Physics*, page 114586, 2025.

Xujia Huang, Fajie Wang, Benrong Zhang, and Hanqing Liu. Enriched physics-informed neural networks for dynamic poisson-nernst-planck systems. *Mathematics and Computers in Simulation*, 237:231–246, 2025.

Arshia Merdasi, Saman Ebrahimi, Xiang Yang, and Robert Kunz. Physics informed neural network application on mixing and heat transfer in combined electroosmotic-pressure driven flow. *Chemical Engineering and Processing-Process Intensification*, 193:109540, 2023.

Shengze Cai, Zhicheng Wang, Lu Lu, Tamer A Zaki, and George Em Karniadakis. Deepm&mnet: Inferring the electroconvection multiphysics fields based on operator approximation by neural networks. *Journal of Computational Physics*, 436:110296, 2021.

Ashwin Mani and Kevin M. Wang. Electroconvection near electrochemical interfaces: Experiments, modeling, and computation. *Annual Review of Fluid Mechanics*, 52(1): 509–529, 2020.

A. Full Final L_2 Results

All values: 3-seed mean $\pm$ std (ddof=1).

A.1. Thermoelasticity — Final L_2

Method	$\gamma=1$	CV	$\gamma=5$	$\gamma=10$	$\gamma=25$	$\gamma=50$	CV	Deg
Adam	4.23e-3±1.41e-3	33%	2.14e-3±8.06e-4	5.78e-3±3.94e-3	9.95e-3±3.33e-3	1.38e-2±5.05e-3	37%	3.3×
SOAP	7.91e-5±1.86e-5	24%	7.71e-5±7.78e-6	8.57e-5±1.11e-5	9.70e-5±1.84e-5	9.83e-5±1.47e-5	15%	1.2×
Adam+LRA	8.88e-4±1.21e-3	136%†	2.83e-3±4.87e-3	3.32e-5±2.35e-5	1.78e-4±7.75e-5	1.70e-4±1.23e-4	72%	0.2×†
SOAP+LRA	2.76e-4±2.35e-4	85%	1.94e-4±1.75e-4	8.52e-5±1.47e-5	3.92e-5±2.13e-5	3.83e-4±2.53e-4	66%	1.4×
Adam+GN	4.34e-3±6.68e-3	154%†	9.41e-3±7.56e-3	3.57e-3±4.53e-3	6.22e-3±4.30e-3	1.40e-3±1.76e-3	126%†	0.3×†
SOAP+GN	6.26e-5±2.15e-5	34%	4.56e-5±2.39e-5	1.18e-4±1.66e-5	1.01e-4±1.10e-4	6.87e-5±8.29e-6	12%	1.1×

CV = coefficient of variation (std/mean $\times$ 100%) at weak- and strong-coupling endpoints. †High CV indicates unreliable degradation ratio.

A.2. Reaction–Diffusion — Final L_2

Method	$k=1$	CV	$k=5$	$k=10$	$k=30$	CV	Deg
Adam	3.47e-3±2.34e-3	67%	5.74e-3±5.62e-3	3.94e-3±1.23e-3	1.12e-2±3.43e-3	31%	3.2×
SOAP	9.90e-5±4.42e-5	45%	1.07e-4±4.56e-5	8.37e-5±3.45e-5	8.05e-5±4.91e-5	61%	0.8×
Adam+LRA	4.44e-5±6.02e-5	136%†	1.64e-4±2.35e-4	1.79e-3±1.46e-3	8.29e-4±4.73e-4	57%	18.7×
SOAP+LRA	7.58e-5±5.61e-5	74%	3.42e-4±2.48e-4	6.40e-5±1.72e-5	1.96e-4±1.35e-4	69%	2.6×
Adam+GN	1.55e-3±1.32e-3	85%†	4.63e-4±5.20e-4	3.40e-3±2.61e-3	7.68e-4±5.51e-4	72%†	0.5×†
SOAP+GN	1.04e-4±3.31e-5	32%	7.48e-5±4.01e-5	6.78e-5±3.10e-5	9.30e-5±1.78e-5	19%	0.9×

A.3. Nernst–Planck–Poisson — Final L_2

Method	$\varepsilon=1.0$	CV	$\varepsilon=0.5$	$\varepsilon=0.2$	CV	Deg
Adam	3.74e-3±5.90e-3	158%	2.09e-3±1.59e-3	3.76e-2±2.33e-3	6%	10×
SOAP	8.53e-5±4.36e-5	51%	2.08e-4±8.85e-5	9.31e-4±2.92e-4	31%	11×
Adam+LRA	4.90e-1±1.39e-2	3%	5.73e-1±4.44e-2	8.62e-1±6.38e-2	7%	1.8×*
SOAP+LRA	2.90e-4±6.79e-5	23%	1.46e-4±1.34e-4	5.25e-1±4.54e-1	86%	1809×*
Adam+GN	3.55e-3±1.61e-3	45%	1.96e-3±1.01e-3	3.65e-1±3.18e-1	87%	103×
SOAP+GN	2.21e-4±1.12e-4	51%	7.89e-5±1.87e-5	2.31e-4±6.15e-5	27%	1.0×

*Adam+LRA: all runs fail ($L_2 > 0.47$). SOAP+LRA: 2/3 seeds fail at $\varepsilon=0.2$.

A.4. 2D NPPS — Final L_2 ($\zeta = 2.0$, 3-seed mean $\pm$ std)

ε	SOAP+GN	Adam+GN
0.2	1.32e-3 ± 8.14e-4	9.12e-1 ± 7.41e-2
0.1	2.85e-3 ± 1.78e-3	1.07e+0 ± 9.76e-3
0.05	1.50e-2 ± 1.07e-2	1.27e+0 ± 4.36e-3

A.5. 2D Adam+GN Training History

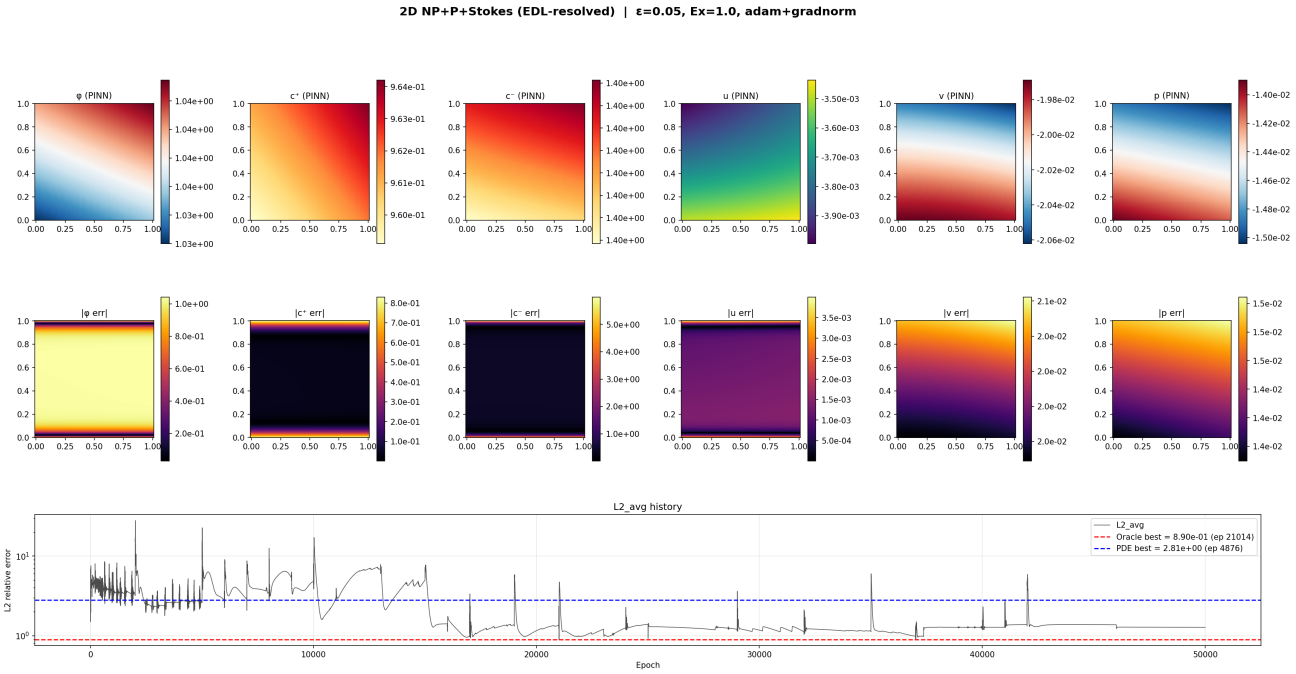


Figure 4. 2D NP+P+Stokes training diagnostics for Adam+GN ($\varepsilon = 0.05$, $\zeta = 2.0$). Top rows: predicted fields and pointwise absolute errors. The electric potential ϕ and ion concentrations $c_\pm$ show $O(1)$ errors across the domain, indicating collapse to a near-constant quasi-neutral state. The Stokes variables (u, v, p) retain $O(10^{-2}$–$10^{-3})$ errors due to the weak coupling from the collapsed ion/potential fields. Bottom: L_2 training history over 50,000 epochs. The optimizer fails to reduce L_2 below 0.9 (oracle-best $L_2 = 0.89$ at epoch 21,014; red dashed line), exhibiting persistent oscillation rather than convergence. This confirms that the $L_2 > 0.9$ failure reported in Table 5 is structural, not due to insufficient training.

A.6. 2D SOAP+GN Training History

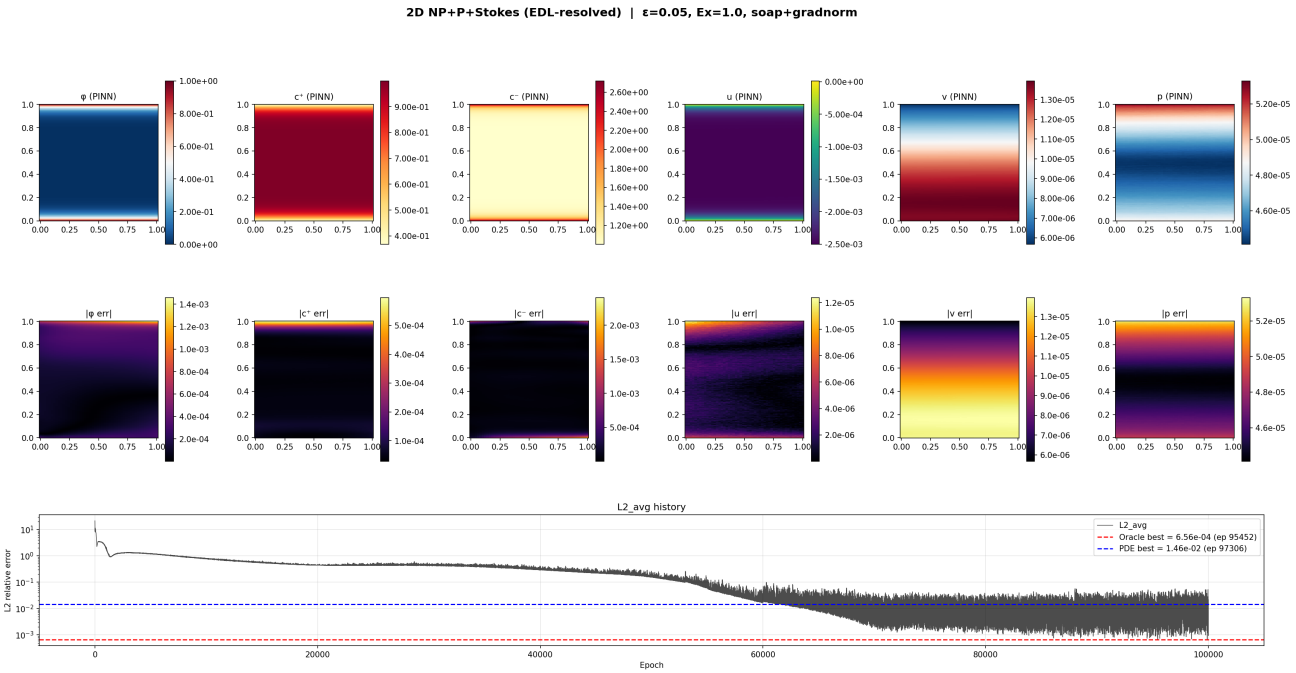


Figure 5. 2D NP+P+Stokes training diagnostics for SOAP+GN ($\varepsilon = 0.05$, $\zeta = 2.0$). Top rows: predicted fields and pointwise absolute errors. The electric potential ϕ and ion concentrations $c_\pm$ show errors of $O(10^{-3}$–$10^{-4})$ concentrated near the Debye layer, indicating successful resolution of the EDL structure. The Stokes variables (u, v, p) achieve $O(10^{-5}$–$10^{-6})$ errors. Bottom: L_2 training history over 100,000 epochs. After an initial transient ($L_2 > 1$ for the first ∼5,000 epochs), the optimizer steadily reduces L_2 to 6.56×10^{-4} (oracle-best at epoch 95,452). Contrast with Adam+GN (Figure 4), which fails to reduce L_2 below 0.9 on the same problem.

B. Theoretical Analysis

We now analyze why coupling strength degrades gradient descent but not Gauss–Newton preconditioning, through the lens of the neural tangent kernel (NTK). While NTK-based convergence analysis for single-equation PINNs has been developed by Wang et al. (2022), the multiphysics setting introduces qualitatively different spectral structure that we characterize here.

B.1. Setup and Notation

Consider S independent networks with disjoint parameters $\theta = (\theta_1, \ldots, \theta_S)$ and M PDE residuals evaluated at N_r collocation points. The residual Jacobian has the block structure

$$J = \left[J_{i,\theta_k}\right]_{\substack{i=1,\ldots,M \\ k=1,\ldots,S}} = \begin{bmatrix} J_{1,\theta_1} & J_{1,\theta_2} & \cdots & J_{1,\theta_S} \\ J_{2,\theta_1} & J_{2,\theta_2} & \cdots & J_{2,\theta_S} \\ \vdots & \vdots & \ddots & \vdots \\ J_{M,\theta_1} & J_{M,\theta_2} & \cdots & J_{M,\theta_S} \end{bmatrix} \in \mathbb{R}^{MN_r \times p}, \tag{6}$$

where each row-block group corresponds to a PDE residual and each column-block group to a network: $J_{i,\theta_k} = \partial \mathbf{R}_i / \partial \theta_k \in \mathbb{R}^{N_r \times p_k}$, $\mathbf{R}_i = (\mathcal{R}_i(x_1), \ldots, \mathcal{R}_i(x_{N_r}))^\top$, and $p = \sum_k p_k$.

The *standard NTK* (relevant for gradient descent) is

$$K = JJ^\top \in \mathbb{R}^{MN_r \times MN_r}. \tag{7}$$

The *block-diagonal Gauss–Newton Hessian* groups by network:

$$H = \mathrm{blkdiag}(H_1, \ldots, H_S), \quad H_k = J_{\theta_k}^\top J_{\theta_k}, \tag{8}$$

where $J_{\theta_k} \in \mathbb{R}^{MN_r \times p_k}$ collects all residual derivatives with respect to θ_k. Note that H retains only the diagonal blocks of the full Gauss–Newton Hessian $J^\top J$; the off-diagonal blocks $J_{\theta_k}^\top J_{\theta_l}$ $(k \neq l)$, which encode cross-network curvature interactions, are discarded. The *preconditioned NTK* is

$$K_P = J\, H^+\, J^\top, \tag{9}$$

where H^+ denotes the Moore–Penrose pseudoinverse.

The linearized residual dynamics $\mathbf{r}_{t+1} = (I - \eta K_P)\mathbf{r}_t$ converge when $\eta < 2/\lambda_{\max}(K_P)$ (Wang et al., 2022). Different choices of preconditioner lead to different K_P and thus different maximum stable learning rates:

Table 6. Preconditioner hierarchy for segregated-network PINNs with S networks, each having L layers and p_k parameters.

Preconditioner	Definition	$\lambda_{\max}(K_P)$	Cost/step	Scalable
None (GD)	$P = I$	$O(\gamma^2)$	$O(p)$	✓
Adam (diagonal)	$P \approx \mathrm{diag}(H)^{-1}$	$O(\gamma^2)$	$O(p)$	✓
Full GN	$P = (J^\top J)^{-1}$	1	$O(p^3)$	×
Block-diag GN	$P = H^{-1}$ (Equation (8))	$\leq S$	$O(\sum p_k^2)$	△
SOAP (Kronecker)	$P \approx \mathrm{blkdiag}(L_k^{(l)} \otimes R_k^{(l)})^{-1/2}$	$\leq SL$	$O(p)$	✓

Scalable: ✓ = $O(p)$ cost, mini-batch compatible; △ = feasible but $O(\sum p_k^2)$ per step; × = intractable for typical PINN sizes.

The key distinction in Table 6 is between *diagonal* and *block* preconditioners. To see why this matters, consider how coupling enters the optimization. The parameter update under preconditioner P is $\theta \leftarrow \theta - \eta\, P\, \nabla\mathcal{L}$. For gradient descent $(P = I)$ and Adam $(P \approx \mathrm{diag}(H)^{-1})$, the preconditioner is diagonal, meaning it scales each parameter using only the corresponding diagonal entry of H while ignoring all off-diagonal entries that encode cross-parameter interactions. However, in coupled systems the Jacobian (6) contains off-diagonal blocks $J_{i,\theta_k} = \gamma\, C_{ik}$ $(i \neq k)$ that couple different networks' parameters through shared PDE residuals. These cross-network blocks contribute $O(\gamma^2)$ terms to the NTK:

$$K_{[i,i]} = \sum_{j=1}^{S} J_{i,\theta_j}\, J_{i,\theta_j}^\top \;\ni\; \gamma^2\, C_{ik}\, C_{ik}^\top, \tag{10}$$

which inflates $\lambda_{\max}(K)$ as $\Omega(\gamma^2)$. The γ^2 growth originates from cross-network Jacobian products that are summed into H_k, contributing to both diagonal and off-diagonal entries of H_k. A diagonal preconditioner retains only the diagonal entries of H_k, discarding the off-diagonal entries that carry the correlation structure among parameters within each network. As a result, Adam's per-parameter scaling cannot fully compensate for the $O(\gamma^2)$ spectral growth.

By contrast, block-diagonal GN (Equation (8)) inverts each $H_k = J_{\theta_k}^\top J_{\theta_k}$ as a whole, capturing the block structure. The preconditioned NTK $K_P = J\, H^+\, J^\top$ then decomposes into projectors (Appendix B.3) whose eigenvalues are bounded by 1 regardless of γ, yielding $\lambda_{\max}(K_P) \leq S$. SOAP approximates this block-diagonal inversion via Kronecker factorization at $O(p)$ cost, making the coupling-independent bound practically achievable.

The following subsections formalize these bounds: Appendix B.2 proves the $\Omega(\gamma^2)$ lower bound on $\lambda_{\max}(K)$, and Appendix B.3 proves the coupling-independent upper bound $\lambda_{\max}(K_P) \leq S$.

B.2. Standard NTK Grows with Coupling

Consider a general linearly coupled PDE system where each residual takes the form

$$\mathcal{R}_i = \mathcal{L}_i[\hat{u}_i] + \gamma\, \mathcal{G}_i[\hat{u}_k] - f_i = 0, \tag{11}$$

where $\mathcal{L}_i$ is the principal differential operator acting on network i's own output, $\mathcal{G}_i$ is a coupling operator acting on network k's output $(k \neq i)$, and γ is the coupling parameter. Differentiating with respect to θ_k yields $J_{i,\theta_k} = \gamma\, C_{ik}$, where $C_{ik} = \partial\mathcal{G}_i[\hat{u}_k]/\partial\theta_k$ is independent of γ.

This structure arises naturally in all linearly coupled systems in this paper: the thermal stress term $\gamma T'$ in thermoelasticity produces $J_{u,\theta_T} = \gamma\, \partial T'/\partial\theta_T$; reaction rates $k\, c_B$ in reaction–diffusion produce $J_{A,\theta_B} = k\, \partial c_B/\partial\theta_B$; and symmetric coupling γv produces $J_{1,\theta_v} = \gamma\, \partial v/\partial\theta_v$.

Theorem 1 (GD spectral bound for linearly coupled systems)**.** *Suppose there exist a residual index i and a network index k such that $J_{i,\theta_k} = \gamma\, C_{ik}$, where $C_{ik} \in \mathbb{R}^{N_r \times p_k}$ is nonzero and independent of γ. Then the standard NTK satisfies*

$$\lambda_{\max}(K) \geq \gamma^2\, \sigma^2_{\max}(C_{ik}), \tag{12}$$

where $\sigma_{\max}(C_{ik})$ is the largest singular value of C_{ik}. Since $\sigma_{\max}(C_{ik})$ depends only on the network architecture and initialization (not γ), the maximum stable learning rate satisfies $\eta^{\mathrm{GD}}_{\max} = O(1/\gamma^2)$.

Proof sketch. The (i,i)-th diagonal block of $K = JJ^\top$ satisfies $K_{[i,i]} = \sum_{j=1}^{S} J_{i,\theta_j} J_{i,\theta_j}^\top \succeq J_{i,\theta_k} J_{i,\theta_k}^\top = \gamma^2 C_{ik} C_{ik}^\top$. Since $K_{[i,i]}$ is a principal submatrix of the PSD matrix K, $\lambda_{\max}(K) \geq \lambda_{\max}(K_{[i,i]}) \geq \gamma^2 \sigma^2_{\max}(C_{ik})$. □

Remark 1 (Scope of Theorem 1)**.** *Theorem 1 applies to any system where coupling enters the Jacobian as $J_{i,\theta_k} = \gamma\, C_{ik}$ with C_{ik} independent of γ. This covers all linearly coupled systems in this paper (thermoelasticity, reaction–diffusion, and the symmetric model used for NTK verification). It does* not *apply to the Nernst–Planck–Poisson system, where coupling is state-dependent ($c_\pm \phi'$): the cross-network Jacobian blocks depend on the current solution rather than a fixed parameter. Accordingly, $\lambda_{\max}(K)$ does not grow with $1/\varepsilon$ at initialization in NP+P (Table 7). The coupling difficulty in NP+P manifests during training as the solution develops boundary-layer structure, but our theory does not provide a formal mechanism for* why *gradient descent fails in this regime—only for why block-diagonal GN preconditioning maintains bounded spectral radius regardless (Theorem 2).*

Remark 2 (Heuristic for nonlinear coupling)**.** *Although Theorem 1 does not formally apply to NP+P, the mechanism can be understood heuristically. The cross-network Jacobian block for the Poisson residual $\varepsilon^2\phi'' + (c_+ - c_-)$ with respect to θ_{c_+} is $J_{\phi,\theta_{c_+}} = \partial c_+/\partial\theta_{c_+}$, which is $O(1)$ and ε-independent at initialization (when the network outputs are near-constant). However, as training progresses and the network develops boundary-layer structure with $c_\pm \sim \exp(\mp\phi)$, the NP residuals $c''_\pm \pm (c_\pm\phi')'$ produce cross-network Jacobian blocks that scale with $|\phi'| = O(1/\varepsilon)$ near the boundary layer. This suggests an* effective *spectral growth of $O(1/\varepsilon^2)$ during training, analogous to the $\Omega(\gamma^2)$ scaling in the linear case—but arising from solution evolution rather than PDE coefficients.*

This heuristic is consistent with two empirical observations: (i) $\lambda_{\max}(K)$ does not grow with $1/\varepsilon$ at initialization (Table 7), confirming that the coupling difficulty is absent before training; (ii) SOAP alone (without GN) degrades $10.9\times$ in NP+P (Table 3), indicating that the coupling difficulty manifests during training. A formal analysis of training-time spectral dynamics for nonlinearly coupled PINNs is left to future work.

B.3. Preconditioned NTK is Coupling-Independent

Recall from Equation (9) that $K_P = J\, H^+\, J^\top$ is the preconditioned NTK under block-diagonal GN. The key insight is that K_P decomposes into a sum of orthogonal projectors.

Lemma 1 (Projector decomposition)**.** *The preconditioned NTK admits the decomposition*

$$K_P = \sum_{k=1}^{S} P_k, \qquad P_k = J_{\theta_k}\, (J_{\theta_k}^\top J_{\theta_k})^+\, J_{\theta_k}^\top, \tag{13}$$

where each P_k is the orthogonal projector onto $\mathrm{col}(J_{\theta_k})$.

Since each P_k has eigenvalues in $\{0, 1\}$, the spectral radius of their sum is bounded:

Theorem 2 (GN spectral bound)**.** *For any segregated-network PINN with S independent networks and block-diagonal GN preconditioning,*

$$\lambda_{\max}(K_P) \leq S. \tag{14}$$

This bound is independent of the coupling parameter γ, the PDE structure, and the network parameterization. The only structural assumptions are disjoint parameters ($\theta_i \cap \theta_j = \emptyset$) and block-diagonal GN.

Proof. From Lemma 1, $K_P = \sum_{k=1}^{S} P_k$. For any unit vector $\mathbf{v}$,

$$\mathbf{v}^\top K_P \, \mathbf{v} = \sum_{k=1}^{S} \|P_k \mathbf{v}\|^2 \leq \sum_{k=1}^{S} 1 = S, \tag{15}$$

where $\|P_k \mathbf{v}\| \leq \|\mathbf{v}\| = 1$ since P_k is an orthogonal projector. □

Remark 3 (Tightness and overparameterized regime)**.** *When $p_k > MN_r$ and J_{θ_k} has full row rank, $P_k = I_{MN_r}$ and $K_P = S \cdot I$, so the bound is achieved with* equality. *For the standard 4-hidden-layer, 64-neuron architecture with $N_r = 200$ ($p_k = 12{,}673 \gg MN_r$), this gives $\lambda_{\max}(K_P) = S$ exactly.*

Remark 4 (Scope and regime-independence of the spectral bounds)**.** *The learning rate bound $\eta_{\max} = 2/\lambda_{\max}(K_P)$ characterizes single-step stability of the linearized residual dynamics $\mathbf{r}_{t+1} = (I - \eta K_P)\mathbf{r}_t$. While this does not constitute a global convergence guarantee for practical (non-lazy) training, a crucial property is that the projector decomposition $K_P = \sum_k P_k$ (Lemma 1) and the resulting bound $\lambda_{\max}(K_P) \leq S$ (Theorem 2) are algebraic identities that hold at* any *parameter configuration θ—they do not require the lazy/NTK regime where θ stays near initialization. This is because $P_k = J_{\theta_k}(J_{\theta_k}^\top J_{\theta_k})^+ J_{\theta_k}^\top$ is an orthogonal projector for any J_{θ_k}, regardless of how θ has evolved during training. The qualitative contrast $\eta_{\max}^{\mathrm{GD}} = O(1/\gamma^2)$ versus $\eta_{\max}^{\mathrm{GN}} = O(1)$ therefore persists throughout training and correctly predicts the empirical observation that Adam degrades with coupling while SOAP+GN remains robust (Section 5).*

Proposition 1 (SOAP's preconditioner is network-wise block-diagonal)**.** *Let each of the S networks have L weight matrices with disjoint parameters ($\theta_i \cap \theta_j = \emptyset$ for $i \neq j$), and let SOAP's layer-wise preconditioner be $H^{\mathrm{SOAP}} = \mathrm{blkdiag}(\tilde{H}_k^{(l)})_{k=1,\ldots,S;\, l=1,\ldots,L}$, where $\tilde{H}_k^{(l)}$ is the Kronecker-factored approximation of $J_{\theta_k^{(l)}}^\top J_{\theta_k^{(l)}}$. Then:*

1. Network-wise block-diagonality. *Since each weight matrix belongs to exactly one network, H^{SOAP} is automatically block-diagonal across networks: $(H^{\mathrm{SOAP}})_{\theta_k,\theta_{k'}} = 0$ for $k \neq k'$.*
2. Coupling-independent spectral bound. *The preconditioned NTK satisfies $\lambda_{\max}(K_P^{\mathrm{SOAP}}) \leq SL$, independent of coupling strength.*
3. Projector ordering. *Since $\mathrm{col}(J_{\theta_k^{(l)}}) \subseteq \mathrm{col}(J_{\theta_k})$ for all l, projection onto the smaller subspace satisfies $\|P_k^{(l)}\mathbf{v}\|^2 \leq \|P_k \mathbf{v}\|^2$ for all $\mathbf{v}$, hence $P_k^{(l)} \preceq P_k$.*

Proof. (1) Immediate from the fact that the layer-level index set $\{(k,l)\}$ is a refinement of the network-level partition $\{k\}$.

(2) From the block-diagonal structure, $K_P^{\mathrm{SOAP}} = \sum_{k=1}^{S}\sum_{l=1}^{L} P_k^{(l)}$, where $P_k^{(l)} = J_{\theta_k^{(l)}}(\tilde{H}_k^{(l)})^+ J_{\theta_k^{(l)}}^\top$ is the orthogonal projector onto $\mathrm{col}(J_{\theta_k^{(l)}})$ (when the Kronecker approximation is exact). By the same argument as Theorem 2: for any unit vector $\mathbf{v}$, $\mathbf{v}^\top K_P^{\mathrm{SOAP}} \mathbf{v} = \sum_{k,l} \|P_k^{(l)}\mathbf{v}\|^2 \leq SL$.

(3) $J_{\theta_k^{(l)}}$ consists of the columns of J_{θ_k} corresponding to layer l; hence $\mathrm{col}(J_{\theta_k^{(l)}}) \subseteq \mathrm{col}(J_{\theta_k})$. For any $\mathbf{v}$, $P_k^{(l)}\mathbf{v}$ is the projection onto a subspace of $\mathrm{col}(J_{\theta_k})$, so $\|P_k^{(l)}\mathbf{v}\|^2 \leq \|P_k \mathbf{v}\|^2$. □

Corollary 1 (Learning rate contrast)**.** *As coupling strengthens ($\gamma \to \infty$):*

$$\eta_{\max}^{\mathrm{GD}} = \frac{2}{\lambda_{\max}(K)} = O(1/\gamma^2) \to 0, \tag{16}$$

$$\eta_{\max}^{\mathrm{GN}} = \frac{2}{\lambda_{\max}(K_P)} = O(1) \geq \frac{2}{S}. \tag{17}$$

C. Proofs

C.1. Proof of Theorem 1

The (i,i)-th diagonal block of $K = JJ^\top$ is $K_{[i,i]} = \sum_{j=1}^{S} J_{i,\theta_j} J_{i,\theta_j}^\top$. Since every term in the sum is positive semidefinite, $K_{[i,i]} \succeq J_{i,\theta_k} J_{i,\theta_k}^\top = \gamma^2 C_{ik} C_{ik}^\top$, hence $\lambda_{\max}(K_{[i,i]}) \geq \gamma^2 \sigma_{\max}^2(C_{ik})$. Since $K_{[i,i]}$ is a principal submatrix of the PSD

matrix $K \in \mathbb{R}^{MN_r \times MN_r}$, $\lambda_{\max}(K) \geq \lambda_{\max}(K_{[i,i]})$ follows from the variational characterization of eigenvalues (any unit vector in the subspace can be zero-padded to a unit vector in the full space with equal Rayleigh quotient), completing the proof.

C.2. Proof of Lemma 1

By the block structure of H, $K_P = JH^+J^\top = \sum_k J_{\theta_k} H_k^+ J_{\theta_k}^\top$. For each term, $J_{\theta_k}(J_{\theta_k}^\top J_{\theta_k})^+ J_{\theta_k}^\top$ is the orthogonal projector onto $\mathrm{col}(J_{\theta_k})$ (standard result from matrix analysis).

C.3. Proof of Theorem 2

From Lemma 1, $K_P = \sum_k P_k$ where each P_k is an orthogonal projector ($P_k^2 = P_k$, eigenvalues in $\{0, 1\}$). For any unit vector $\mathbf{v}$ ($\|\mathbf{v}\| = 1$):

$$\mathbf{v}^\top K_P \, \mathbf{v} = \sum_{k=1}^{S} \mathbf{v}^\top P_k \mathbf{v} = \sum_{k=1}^{S} \|P_k \mathbf{v}\|^2.$$

Since each P_k is an orthogonal projector, $\|P_k \mathbf{v}\| \leq \|\mathbf{v}\| = 1$. Summing over S networks gives $\mathbf{v}^\top K_P \mathbf{v} \leq S$ for all unit vectors, hence $\lambda_{\max}(K_P) \leq S$.

Equality condition. If $\mathbf{v} \in \mathrm{col}(J_{\theta_k})$ for all k, then $P_k \mathbf{v} = \mathbf{v}$ and $\|P_k \mathbf{v}\| = 1$ for all k, achieving $\mathbf{v}^\top K_P \mathbf{v} = S$. In the overparameterized regime ($p_k > MN_r$ with J_{θ_k} full row rank), $\mathrm{col}(J_{\theta_k}) = \mathbb{R}^{MN_r}$ for all k, so $P_k = I$ and $K_P = SI$.

Note on generality. This proof uses no properties of the PDE residuals, coupling structure, or network architecture beyond the disjoint-parameter assumption (Equation (8)). The bound $\lambda_{\max}(K_P) \leq S$ holds for any segregated-network PINN regardless of coupling type (linear, nonlinear, one-way, bidirectional), PDE coefficients, or network width/depth.

C.4. Proof of Proposition 1

Part (1). The layer-level index set $\{(k, l) : k = 1, \ldots, S,\ l = 1, \ldots, L\}$ is a refinement of the network-level partition $\{k : k = 1, \ldots, S\}$: each network group $\{\theta_k\}$ equals the union $\bigcup_{l=1}^{L} \{\theta_k^{(l)}\}$. A matrix that is block-diagonal with respect to a finer partition is block-diagonal with respect to any coarser partition. Since H^{SOAP} has zero blocks between any two distinct index pairs $(k, l) \neq (k', l')$, it has zero blocks between distinct networks $k \neq k'$.

Part (2). From the block-diagonal structure, $K_P^{\mathrm{SOAP}} = \sum_{k,l} P_k^{(l)}$ where $P_k^{(l)} = J_{\theta_k^{(l)}} (\tilde{H}_k^{(l)})^+ J_{\theta_k^{(l)}}^\top$ is the orthogonal projector onto $\mathrm{col}(J_{\theta_k^{(l)}})$ (when the Kronecker approximation is exact). For any unit vector $\mathbf{v}$:

$$\mathbf{v}^\top K_P^{\mathrm{SOAP}} \, \mathbf{v} = \sum_{k=1}^{S} \sum_{l=1}^{L} \|P_k^{(l)} \mathbf{v}\|^2 \leq \sum_{k=1}^{S} \sum_{l=1}^{L} 1 = SL.$$

Part (3). $J_{\theta_k^{(l)}}$ consists of the columns of J_{θ_k} corresponding to the parameters of layer l; hence $\mathrm{col}(J_{\theta_k^{(l)}}) \subseteq \mathrm{col}(J_{\theta_k})$. For any vector $\mathbf{v}$, $P_k^{(l)} \mathbf{v}$ is the orthogonal projection onto a subspace of $\mathrm{col}(J_{\theta_k})$, so $\|P_k^{(l)} \mathbf{v}\|^2 \leq \|P_k \mathbf{v}\|^2$, i.e., $P_k^{(l)} \preceq P_k$.

D. NTK Numerical Verification

We verify the spectral predictions of Appendix B by computing $\lambda_{\max}(K)$ and $\lambda_{\max}(K_P)$ at network initialization across all benchmark systems.

D.1. Methodology

For each system, we compute the full Jacobian J via per-sample backpropagation and form $K = JJ^\top$. The preconditioned NTK is computed via the projector decomposition (13): for each network k, we compute the economy SVD of $J_{\theta_k} = U_k \Sigma_k V_k^\top$ and form $P_k = U_k U_k^\top$ (retaining singular vectors with $\sigma > 10^{-10} \sigma_{\max}$). This avoids explicit inversion of H_k, which has condition number $\sim 10^{18}$ for typical MLP architectures.

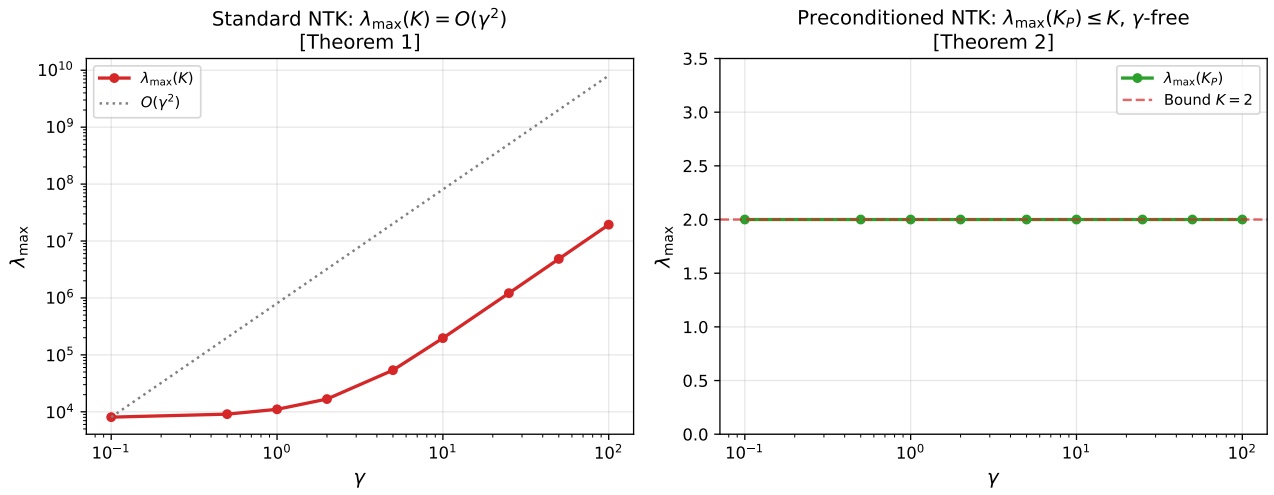


Figure 6. NTK spectral analysis at initialization for the symmetric coupled system. **(a)** $\lambda_{\max}(K)$ grows as $\Omega(\gamma^2)$ (dashed line), confirming Theorem 1. **(b)** $\lambda_{\max}(K_P) = 2.0 = S$ across all γ and seeds, confirming Theorem 2 with equality (Remark 3).

D.2. Symmetric Model Problem

We first verify on a symmetric model problem: $u_{xx}+\gamma v = f_1$, $v_{xx}+\gamma u = f_2$ on $[0, 1]$ with exact solutions $u^* = \sin(\pi x)+1$, $v^* = \sin(2\pi x)$. Both cross-network blocks satisfy the linear coupling assumption of Theorem 1 with $C_{1v} = \partial v/\partial\theta_v$ and $C_{2u} = \partial u/\partial\theta_u$. Using the same architecture as the benchmarks (4 hidden layers, 64 neurons, $N_r = 200$), we sweep $\gamma \in \{0.1, 0.5, 1, 2, 5, 10, 25, 50, 100\}$ with 3 seeds each.

Figure 6 shows the results. $\lambda_{\max}(K)$ grows as $\Omega(\gamma^2)$ in the coupling-dominated regime ($\gamma \geq 10$, log-log slope $= 2.0$), while $\lambda_{\max}(K_P) = 2.000$ across all 27 experiments, confirming Theorems 1 and 2.

D.3. Training Dynamics

Theorems 1 and 2 hold at any parameter configuration θ, not only at initialization. We verify this by measuring K and K_P at 8 snapshots during 30,000 epochs of training (Figure 7). While $\lambda_{\max}(K)$ varies by two orders of magnitude during training (confirming that the network is *not* in the lazy/NTK regime), $\lambda_{\max}(K_P) = 2.000$ at every snapshot, for both Adam and SOAP+GN trajectories, across $\gamma \in \{1, 10, 50\}$. This empirically confirms the regime-independence argued in Remark 4: the projector bound is an algebraic identity that does not rely on parameter proximity to initialization.

D.4. Generalization to Benchmark Systems

Table 7 shows the spectral measurements across all benchmark systems. Thermoelasticity and reaction–diffusion satisfy the linear coupling assumption of Theorem 1 directly (see Remark 1), while NP+P has state-dependent coupling where Theorem 1 does not apply.

Two observations: (1) $\lambda_{\max}(K_P)$ equals the bound S with equality in all systems, confirming Theorem 2 and the tightness predicted by Remark 3 for the overparameterized regime. (2) In NP+P, $\lambda_{\max}(K)$ does not grow with $1/\varepsilon$ at initialization, consistent with the scope limitation of Theorem 1 noted in Remark 1: the coupling is state-dependent ($c_{\pm}\phi'$) rather than parametrically linear. The coupling difficulty in NP+P manifests during training rather than at initialization, consistent with the observation that SOAP alone (without GN) degrades $10.9\times$ in this system (Table 3).

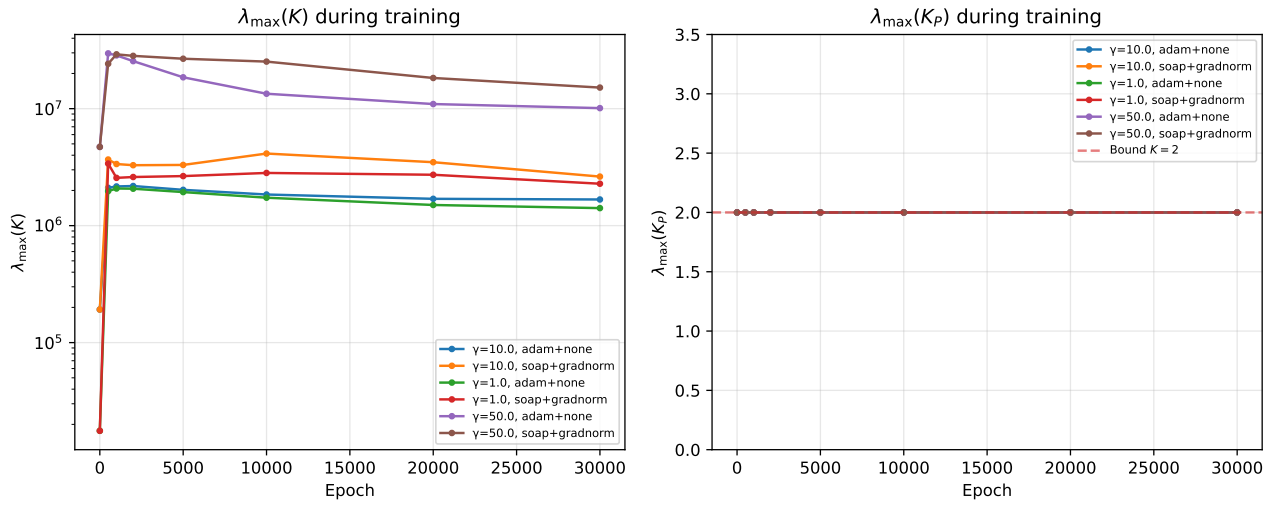


Figure 7. NTK spectral quantities during training. **(a)** $\lambda_{\max}(K)$ varies by $10^2\times$ during training, confirming non-lazy regime. **(b)** $\lambda_{\max}(K_P) = 2.0$ at every snapshot, every optimizer, every γ—the bound holds throughout training, not only at initialization.

Table 7. NTK spectral measurements at initialization across all benchmark systems (3-seed or 2-seed mean). S = number of networks. In all cases, $\lambda_{\max}(K_P) = S$ (overparameterized regime), independent of coupling strength.

System	Coupling param.	S	$\lambda_{\max}(K)$	$\lambda_{\max}(K_P)$	Bound S
Symmetric	$\gamma = 0.1$	2	8.1e3	**2.0**	2
(theory model)	$\gamma = 100$	2	1.9e7	**2.0**	2
Thermoelasticity	$\gamma = 1$	2	1.4e4	**2.0**	2
(asymmetric)	$\gamma = 50$	2	1.8e7	**2.0**	2
Reaction–Diff.	$k = 0.5$	3	1.8e4	**3.0**	3
(3-species)	$k = 30$	3	1.5e7	**3.0**	3
NP+P	$\varepsilon = 1.0$	3	2.2e4	**3.0**	3
(nonlinear)	$\varepsilon = 0.1$	3	1.9e4	**3.0**	3