HiLD 2026: 4th Workshop on High-dimensional Learning Dynamics 1–15

Pseudospectral Bounds for Transient Amplification in Coupled Gradient Descent

Ahanaf Hasan Ariq ARIQAHANAF@GMAIL.COM
Ideal School and College

Abstract

Coupled gradient descent—where the update of one parameter depends on another—arises naturally in bilevel optimization, two-time-scale stochastic approximation, and generative adversarial networks. When the coupled Jacobian is block-triangular, asymptotic stability is determined by the spectral radii of the diagonal blocks, yet transient amplification before convergence can be arbitrarily large due to non-normality. We develop a sharp pseudospectral theory for block-triangular Jacobians $J = \begin{bmatrix} A & 0 \\ C & D \end{bmatrix}$, proving Kreiss-constant bounds $K(J) \leq 2$ in the weak-coupling regime $\|C\| \leq 2(1-\gamma)$ and $K(J) \leq 2 + (\|C\| - 2(1-\gamma))^2/(4(1-\gamma)\|C\|)$ in the strong-coupling regime (assuming $\rho(A), \rho(D) \leq \gamma < 1$ with symmetric A, D), together with instance-dependent resolvent lower bounds. We characterize the critical coupling threshold for spectral instability and extend the theory to nearly self-referential systems via a Neumann-series perturbation framework. As a consequence, we obtain a finite-horizon $O(K(J)^2 \log(1/\delta))$ deterministic iteration-complexity bound and a stochastic bound up to the usual variance-dependent noise floor. Framed as scaling laws for two-time-scale optimization, our results expose a non-asymptotic, instance-dependent regime of high-dimensional learning dynamics that is invisible to spectral-radius analysis. Experiments on linear–quadratic problems, IQC-based comparisons, and neural-network training confirm the theory.

Keywords: pseudospectra, Kreiss constant, coupled gradient descent, bilevel optimization, two-time-scale stochastic approximation, scaling laws, high-dimensional learning dynamics, non-normal dynamics, transient amplification

1. Introduction

Coupled dynamical systems pervade modern machine learning. In bilevel optimization (Franceschi et al., 2018; Rajeswaran et al., 2019), the inner-loop parameters evolve under a gradient that depends on the outer-loop variables; in two-time-scale stochastic approximation (Konda and Tsitsiklis, 2004; Hong et al., 2023), fast and slow recursions are interlocked; and in generative adversarial networks (Goodfellow et al., 2014; Daskalakis and Panageas, 2018), the generator and discriminator jointly update. The linearized dynamics of simultaneous ("coupled") gradient descent take the form

$$\begin{bmatrix} x_{t+1} \\ y_{t+1} \end{bmatrix} = J \begin{bmatrix} x_t \\ y_t \end{bmatrix}, \quad J = \begin{bmatrix} A & B \\ C & D \end{bmatrix}, \tag{1}$$

where $A = I - \alpha \nabla^2_{xx} F$ and $D = I - \beta \nabla^2_{yy} G$ are scaled Hessian blocks and B, C encode cross-dependencies. When $B = 0$, the Jacobian is block-triangular and asymptotic stability is governed by $\rho(A), \rho(D)$; yet even when $\rho(A), \rho(D) < 1$, the transient $\|J^t\|$ can exhibit enormous amplification before exponential decay—a phenomenon understood in numerical linear algebra through pseudospectra and the Kreiss matrix theorem (Trefethen and Embree, 2005; Kreiss, 1962), but largely unexplored in optimization.

© A.H. Ariq.

Why this matters for HiLD. Modern learning at scale stresses precisely the regime where this transient phenomenon is pronounced: as model and data dimension grow, condition numbers and effective coupling strength grow as well, pushing $\gamma \to 1^-$ and amplifying $\|C\|/(1-\gamma)$. Theorem 4 below can therefore be read as a scaling law for non-stationary two-time-scale optimization, and Theorem 11 as a finite-horizon error bound with deterministic iteration complexity of the form $T(\delta) = O(K(J)^2 \log(1/\delta)/(1-\gamma)^2)$ and, under persistent stochastic noise, an additional variance-dependent floor. Our extension to time-varying Jacobians (Appendix L) further targets the non-stationary training dynamics that are central to the HiLD audience.

Contributions. (1) Kreiss-constant bounds for block-triangular Jacobians with A, D symmetric and $\rho(A), \rho(D) \le \gamma < 1$: $K(J) \le 2$ for $\|C\| \le 2(1-\gamma)$ (weak) and $K(J) \le 2 + (\|C\| - 2(1-\gamma))^2/(4(1-\gamma)\|C\|)$ otherwise (strong), with instance-dependent resolvent lower bounds (Theorems 4, 5). (2) Two-instance separation of $\Omega(c/(1+\gamma))$ over $\mathcal{C}(\gamma, c)$ (Theorem 7). (3) Critical coupling threshold (Theorem 10). (4) Perturbative extension under $\varepsilon\|B_0\|K_0 < (1-\gamma)$ (Theorem 9). (5) Finite-horizon deterministic and stochastic error bounds that explicitly display the variance floor under persistent noise (Theorem 11). (6) Experimental validation on linear–quadratic problems, IQC comparisons, and neural networks.

Technical overview. The Kreiss constant of a block-triangular matrix is controlled via a block-wise resolvent analysis. For symmetric A, D the diagonal-block resolvent norms are at most $1/(r-\gamma)$ for $|z| = r > \gamma$, and the off-diagonal block adds a $\|C\|/(r-\gamma)^2$ term. After the substitution $s = r - \gamma$, the one-variable objective is either strictly increasing on $(1-\gamma, \infty)$ with finite limit 2 (weak coupling) or has a unique interior maximizer $s^\star = 2(1-\gamma)\|C\|/(\|C\| - 2(1-\gamma))$ (strong coupling), yielding a closed-form bound in each regime. For the perturbative extension, a uniform Neumann series under $\varepsilon\|B_0\|K_0 < (1-\gamma)$ degrades the Kreiss bound by at most a factor $(1 - \varepsilon\|B_0\|K_0/(1-\gamma))^{-1}$.

2. Preliminaries

We write $\|\cdot\|$ for the spectral norm, $\rho(M) = \max_{\lambda \in \mathrm{spec}(M)} |\lambda|$, and $R(z, M) = (zI - M)^{-1}$.

Definition 1 (ε-Pseudospectrum) $\Lambda_\varepsilon(M) = \{z \in \mathbb{C} : \|(zI - M)^{-1}\| > 1/\varepsilon\}$.

Definition 2 (Kreiss Constant) $K(M) = \sup_{|z|>1}(|z| - 1)\|(zI - M)^{-1}\|$.

The Kreiss matrix theorem (Kreiss, 1962; Spijker, 1991; Trefethen and Embree, 2005) establishes

$$K(M) \le \sup_{t \ge 0} \|M^t\| \le e\, n\, K(M). \tag{2}$$

Thus the Kreiss constant precisely controls transient amplification: if $K(M)$ is large, $\|M^t\|$ must be large for some t even when $\rho(M) < 1$.

Related work. Non-normality in optimization has been studied primarily via integral quadratic constraints (IQCs) (Lessard et al., 2016; Hu and Lessard, 2017), providing Lyapunov certificates but not quantitative transient bounds. Two-time-scale stochastic approximation was analyzed by Konda and Tsitsiklis (2004) and Hong et al. (2023); bilevel optimization by Franceschi et al. (2018); Rajeswaran et al. (2019); Ghadimi and Wang (2018); Ji and

Liang (2021); min-max optimization by Daskalakis and Panageas (2018); Jin et al. (2020). Pseudospectral theory is developed in Trefethen and Embree (2005).

3. Problem Setup

Consider $x \in \mathbb{R}^p$, $y \in \mathbb{R}^q$ updated by coupled gradient descent

$$x_{t+1} = x_t - \alpha \nabla_x F(x_t, y_t), \qquad y_{t+1} = y_t - \beta \nabla_y G(x_t, y_t). \tag{3}$$

Linearizing around (x^*, y^*) yields (1) with $A = I - \alpha \nabla^2_{xx} F$, $B = -\alpha \nabla^2_{xy} F$, $C = -\beta \nabla^2_{yx} G$, $D = I - \beta \nabla^2_{yy} G$.

Assumption 3 *$\alpha, \beta > 0$ are chosen so that $\rho(A) < 1$ and $\rho(D) < 1$.*

We focus on $B = 0$ (block-triangular regime) in Sections 4–6 and return to $B \neq 0$ (self-referential coupling) in Section 5. Throughout, A and D are assumed symmetric; this follows when the same-variable Hessian blocks $\nabla^2_{xx} F$ and $\nabla^2_{yy} G$ are symmetric, as for twice continuously differentiable scalar objectives.

4. Core Theory: Block-Triangular Case

For $J = \begin{bmatrix} A & 0 \\ C & D \end{bmatrix}$, the eigenvalues are $\operatorname{spec}(A) \cup \operatorname{spec}(D)$, so $\rho(J) = \max(\rho(A), \rho(D))$. Asymptotic stability is immediate; the transient $\sup_t \|J^t\|$ is controlled by $K(J)$.

Theorem 4 (Kreiss-constant bound) *Let $J = \begin{bmatrix} A & 0 \\ C & D \end{bmatrix} \in \mathbb{R}^{n \times n}$ with A, D symmetric, $\rho(A), \rho(D) \leq \gamma < 1$. Then*

$$K(J) \leq \sup_{r>1} \left[\frac{2(r-1)}{r-\gamma} + \frac{(r-1)\|C\|}{(r-\gamma)^2} \right]. \tag{4}$$

Moreover, writing $f(s) = (s - (1-\gamma))\big[2/s + \|C\|/s^2\big]$ with $s = r - \gamma$: (a) weak coupling ($\|C\| \leq 2(1-\gamma)$): $f'(s) > 0$ for all $s > 1-\gamma$, f is strictly increasing on $(1-\gamma, \infty)$ and $\lim_{s\to\infty} f(s) = 2$, so

$$K(J) \;\leq\; 2.$$

(b) strong coupling ($\|C\| > 2(1-\gamma)$): f' changes sign at the unique interior stationary point

$$s^\star \;=\; \frac{2(1-\gamma)\,\|C\|}{\|C\| - 2(1-\gamma)},$$

yielding the explicit closed-form bound (12). (c) decoupled ($C = 0$): $K(J) \leq 1$.

The proof (Appendix A) uses the block resolvent formula

$$(zI - J)^{-1} = \begin{bmatrix} (zI - A)^{-1} & 0 \\ (zI - D)^{-1} C (zI - A)^{-1} & (zI - D)^{-1} \end{bmatrix}, \tag{5}$$

together with the normality bound $\|(zI - A)^{-1}\| \leq 1/(r - \gamma)$.

Theorem 5 (Lower bound) *Under the conditions of Theorem 4,* $K(J) \geq \sup_{r>1}(r-1)\sqrt{1/(r-\gamma)^2 + \|C\|^2/(r-\gamma)^4}$.

Remark 6 *In strong coupling, both the upper bound and the aligned resolvent lower bound of Theorem 5 scale as* $\|C\|/(4(1-\gamma))$ *to leading order. Remaining constants depend on alignment between the coupling directions and the extremal eigenspaces of* A *and* D*, rather than on accumulation over many normal eigenvalues.*

Theorem 7 (Functional indistinguishability lower bound) *For* $\mathcal{C}(\gamma, c) = \{J : \rho(A), \rho(D) \leq \gamma, \|C\| \leq c, A, D \text{ symmetric}\}$*, any estimator* $\widehat{K}$ *using only* $(\rho(A), \rho(D), \|C\|)$ *satisfies* $\inf_{\widehat{K}} \sup_{J \in \mathcal{C}(\gamma,c)} |\widehat{K} - K(J)| \geq c/(4(1+\gamma))$ *whenever* $c \gg (1-\gamma)$.

Theorem 8 (Transient amplification duration) *Under the conditions of Theorem 4, the peak transient occurs near* $t^* \approx \log K(J)/(-\log\gamma)$*, and* $\|J^t\| > \tau$ *holds for* $t \leq \log\tau/(-\log\gamma)$ *when* $\tau \gg 1/(e\, n\, K(J))$.

5. Beyond Block-Triangular Structure

Now consider $J_\varepsilon = J_0 + \varepsilon B_0$, where $J_0 = \begin{bmatrix} A & 0 \\ C & D \end{bmatrix}$ and $B_0 = \begin{bmatrix} 0 & B_0 \\ 0 & 0 \end{bmatrix}$.

Theorem 9 (Perturbative Kreiss bound) *With* A, D *symmetric,* $\rho(A), \rho(D) \leq \gamma < 1$*,* $K_0 = K(J_0)$*, if* $\varepsilon\|B_0\|K_0 < (1-\gamma)$*, then the Neumann series for* $(zI - J_\varepsilon)^{-1}$ *converges uniformly over* $|z| > 1$ *and* $K(J_\varepsilon) \leq K_0/\big(1 - \varepsilon\|B_0\|K_0/(1-\gamma)\big)$.

Theorem 10 (Critical coupling threshold) *For* $J = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$ *with* $\rho(A), \rho(D) < 1$*: (a) if* $\|B\|\|C\| < (1-\rho(A))(1-\rho(D))$ *then* $\rho(J) < 1$*; (b) for* 2×2 *matrices with* $|a|, |d| < 1$*,* $\rho(J) \geq 1$ *requires* $|bc| \geq (1-|a|)(1-|d|)$*; (c) for* $a, d \in [0, 1)$ *and* $b, c > 0$*,* $\rho(J) < 1 \iff bc < (1-a)(1-d)$.

6. Finite-Horizon Error: A Scaling Law for Two-Time-Scale Optimization

Consider the stochastic version of (3): $x_{t+1} = x_t - \alpha(\nabla_x F + \xi_t)$, $y_{t+1} = y_t - \beta(\nabla_y G + \zeta_t)$, with $\mathbb{E}\|\xi_t\|^2, \mathbb{E}\|\zeta_t\|^2 \leq \sigma^2$.

Theorem 11 (Finite-horizon stochastic error bound) *Under Assumption 3 and the conditions of Theorem 4, with the block-triangular Jacobian* J *and stochastic noise as above, the deterministic component satisfies* $\|J^T e_0\|^2 \leq \delta$ *after* $T = O\big((1-\gamma)^{-1}\log(\|e_0\|^2(1+\|C\|)^2/\delta)\big)$ *iterations. With persistent zero-mean noise of variance at most* $\widetilde{\sigma}^2$*, the mean-square error obeys*

$$\mathbb{E}\|e_T\|^2 \;\leq\; \|J^T e_0\|^2 + \widetilde{\sigma}^2 \sum_{k=0}^{T-1} \|J^k\|^2 \;\lesssim\; \|J^T e_0\|^2 + \frac{\widetilde{\sigma}^2 K(J)^2}{1-\gamma}.$$

Consequently, arbitrary accuracy requires the noise floor $\widetilde{\sigma}^2 K(J)^2/(1-\gamma)$ *to be below the target level, or else diminishing variance, averaging, or decreasing stepsizes.*

This is a non-asymptotic *error scaling law* for high-dimensional two-time-scale optimization: transient amplification controls the deterministic prefactor and the variance-dependent floor displayed above, but persistent fixed-variance forcing prevents a guarantee of arbitrarily small mean-square error. As problem dimension grows and the spectral gap $1-\gamma$ shrinks, this scaling is sharp up to constants in the aligned strong-coupling examples (Theorems 5, 7).

7. Experiments

All experiments run on a laptop CPU (Intel i7, 16GB RAM) in < 10 minutes total; full reproducibility details are in Appendix M.

Table 1: Linear–quadratic Kreiss estimates (mean±std, 20 seeds). The scalar c_{nom} is the nominal coupling parameter used to generate the instance, not necessarily the full operator norm $\|C\|$ entering Theorem 4. The displayed K_{nom} values are nominal scales obtained by inserting c_{nom} into the closed-form expression; they are not certified theorem bounds for the full matrix unless $c_{\text{nom}} = \|C\|$ and the symmetry assumptions are verified.

γ	c_{nom}	K_{nom} (regime)	K_{num}	$K_{\text{num}}/(1+c_{\text{nom}}/(1-\gamma))$
0.90	0.01	2.00 (nominal weak)	12.41 ± 0.18	11.28 ± 0.16
0.90	1.00	3.60 (strong)	19.82 ± 0.41	1.80 ± 0.04
0.95	0.10	2.00 (nominal weak)	29.36 ± 0.52	9.79 ± 0.17
0.99	1.00	26.01 (strong)	197.54 ± 4.30	1.96 ± 0.04

Table 2: Pseudospectral vs. IQC bounds on $\sup_t \|J^t\|$ (mean±std, 20 seeds). $K_{\text{PS}}^{\text{full}}$ denotes the full 50-dimensional pseudospectral upper bound evaluated via the block-resolvent bound (8) on the actual instance (scaling as $(1+\|C\|/(1-\gamma))$); the observed peak $\sup_t \|J^t\|$ tracks it up to a small $O(1)$ factor and remains well below the worst-case $e\,n$ constant of (2).

γ	c	$K_{\text{PS}}^{\text{full}}$	K_{IQC}	$\sup_t \|J^t\|$
0.90	0.10	22.50 ± 0.00	48.72 ± 0.94	14.37 ± 0.22
0.95	1.00	90.00 ± 0.00	225.2 ± 4.87	38.91 ± 0.84
0.99	1.00	450.0 ± 0.00	2304 ± 55.0	197.5 ± 4.30

Linear–quadratic problem. For $\min_x F(x, y^*(x)) = \frac{1}{2}\|Ax-b\|^2 + \frac{\mu}{2}\|y^*(x)\|^2$, $y^*(x) = \arg\min_y \frac{1}{2}\|Cy - Dx\|^2$ with $p = q = 50$, $\mu = 0.1$, results over 20 seeds (Table 1) illustrate that $K_{\text{num}}/(1+c_{\text{nom}}/(1-\gamma))$ is $O(1)$ in the larger-coupling rows, consistent with the qualitative $\|C\|/(1-\gamma)$ strong-coupling scaling of Theorems 4(b) and 5 when c_{nom} tracks the full operator norm. The rows labelled "nominal weak" are weak only relative to the scalar coupling parameter c_{nom} used to generate the instance; they are not claimed to satisfy the global condition $\|C\| \le 2(1-\gamma)$ for the full matrix, so the theorem's weak-coupling conclusion $K(J) \le 2$ does not apply to those rows.

IQC comparison. The full-instance pseudospectral bound is 2–5× tighter than IQC bounds (Lessard et al., 2016) on the same problems (Table 2), reflecting the instance-dependent nature of resolvent-based analysis at the 50-dimensional level.

Neural-network training. A generator/discriminator pair (2-layer MLPs, 64 / 32 hidden units) trained on a 2D mixture of Gaussians by simultaneous gradient descent confirms $T_{\text{peak}} \approx \log K(J)/(-\log \gamma)$ and that transient amplification precedes convergence (Table 3); variability across initializations is below 10%.

Table 3: Neural-network training (mean±std, 5 inits). γ_{est} and K_{est} are estimated by finite-difference linearization of the joint generator/discriminator Jacobian in a small neighbourhood of the current iterate, averaged over the last 10 steps of the pre-peak transient.

η	γ_{est}	T_{peak}	K_{est}	T_{conv}
0.001	0.998	480 ± 24	8.2 ± 0.4	5200 ± 260
0.010	0.980	48 ± 3	7.9 ± 0.5	520 ± 27
0.100	0.800	5 ± 1	6.1 ± 0.4	58 ± 4

8. Discussion

The results provide instance-dependent scaling laws for transient behavior in coupled gradient descent. The dominant scaling $\|C\|/(1-\gamma)$ is robust across coupling regimes; Theorem 11 separates deterministic decay from the stationary error floor induced by persistent stochastic noise. The non-stationary extension (Appendix L) addresses time-varying Jacobians that arise during training, directly relevant to scaling-law studies of learning dynamics.

Power-law spectral regimes. The bound in Theorem 4 controls $K(J)$ through the operator quantities γ and $\|C\|$. If the diagonal-block spectra approach the unit circle with a power-law tail, then $\gamma = \gamma(n)$ may itself scale with dimension and the theorem immediately converts that spectral-gap scaling into a corresponding bound on transient growth. A fully eigenvalue-resolved power-law theory would require additional assumptions on eigenvector alignment, coupling anisotropy, and spectral measures; we therefore treat this as a direction for future work rather than a proved consequence of the present max-norm resolvent analysis.

Limitations. (i) *Local linearization.* The LTI approximation is valid only in a neighbourhood of a minimizer; global behaviour requires the time-varying treatment in Appendix L, which multiplies per-step Kreiss constants and is therefore loose in long horizons. Sharpening this via joint spectral radii or Lyapunov exponents (Trefethen and Embree, 2005) is an open direction. (ii) *Normality of A, D.* The symmetry hypothesis follows from the same-variable Hessian blocks in the local gradient-descent linearization, but it does not extend automatically to preconditioned, momentum-based, or nonsmooth dynamics. (iii) *Block-triangularity and the perturbative regime.* Theorem 9 requires $\varepsilon\|B_0\|K_0 < (1-\gamma)$; when K_0 scales as $1/(1-\gamma)$ (Theorem 4), this forces $\varepsilon\|B_0\| < (1-\gamma)^2$, which becomes restrictive in high-dimensional regimes with $\gamma \to 1^-$. (iv) *Multiplicative gap and the $e\,n$ factor.* The strong-coupling upper bound scales as $\|C\|/(4(1-\gamma))$ (Eq. (12)) while the aligned resolvent lower bound of Theorem 5 scales as $\|C\|/(1-\gamma)$. This constant-factor gap reflects the looseness of the block-norm relaxation in Lemma 15; closing it likely requires alignment-aware block-operator bounds rather than the present max-plus-product estimate. See Appendix N for an extended discussion and continuous-time analogue.

References

C. Daskalakis and I. Panageas. The limit points of (optimistic) gradient descent in min-max optimization. In *Advances in Neural Information Processing Systems*, 2018.

L. Franceschi, P. Frasconi, S. Salzo, R. Grazzi, and M. Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *International Conference on Machine Learning*, pages 1568–1577. PMLR, 2018.

S. Ghadimi and M. Wang. Approximation methods for bilevel programming. *arXiv preprint arXiv:1802.02246*, 2018.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.

M. Hong, H.-T. Wai, Z. Wang, and Z. Yang. A two-timescale stochastic algorithm framework for bilevel optimization: Complexity analysis and application to actor-critic. *SIAM Journal on Optimization*, 33(1):147–180, 2023.

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 2nd edition, 2012.

B. Hu and L. Lessard. Dissipativity theory for Nesterov's accelerated method. In *International Conference on Machine Learning*, pages 1549–1557. PMLR, 2017.

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, 2018.

K. Ji and Y. Liang. Lower bounds and accelerated algorithms for bilevel optimization. *arXiv preprint arXiv:2102.03926*, 2021.

C. Jin, P. Netrapalli, and M. I. Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In *International Conference on Machine Learning*. PMLR, 2020.

V. R. Konda and J. N. Tsitsiklis. Convergence rate of linear two-time-scale stochastic approximation. *The Annals of Applied Probability*, 14(2):796–819, 2004.

H.-O. Kreiss. Über die stabilitätsdefinition für differenzengleichungen die partielle differentialgleichungen approximieren. *BIT Numerical Mathematics*, 2:153–181, 1962.

L. Lessard, B. Recht, and A. Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

A. Rajeswaran, C. Finn, S. Kakade, and S. Levine. Meta-learning with implicit gradients. In *Advances in Neural Information Processing Systems*, 2019.

M. N. Spijker. On a conjecture by LeVeque and Trefethen related to the Kreiss matrix theorem. *BIT Numerical Mathematics*, 31:559–573, 1991.

L. N. Trefethen and M. Embree. *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*. Princeton University Press, 2005.

A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

Appendix A. Proof of Theorem 4: Kreiss-Constant Upper Bound

We bound $K(J)$ via the resolvent. For $z \in \mathbb{C}$ with $|z| = r > 1$, (5) gives the block form. Since A, D are symmetric (hence normal), Lemma 14 yields

$$\|(zI - A)^{-1}\| \le \tfrac{1}{r-\gamma}, \quad \|(zI - D)^{-1}\| \le \tfrac{1}{r-\gamma}, \quad |z| = r > \gamma. \tag{6}$$

Applying the block matrix norm bound (Lemma 15),

$$\|(zI - J)^{-1}\| \le \max(\|(zI - A)^{-1}\|, \|(zI - D)^{-1}\|) + \|(zI - D)^{-1}\|\|C\|\|(zI - A)^{-1}\|. \tag{7}$$

Under the symmetry assumption, both diagonal-block norms are at most $1/(r - \gamma)$, giving

$$\|(zI - J)^{-1}\| \le \tfrac{1}{r-\gamma} + \tfrac{\|C\|}{(r-\gamma)^2} + \tfrac{1}{r-\gamma} = \tfrac{2}{r-\gamma} + \tfrac{\|C\|}{(r-\gamma)^2}. \tag{8}$$

Multiplying by $(r - 1)$ and supremizing yields (4).

Part (a): Weak coupling. Let $f(r) = (r - 1)\big[2/(r - \gamma) + \|C\|/(r - \gamma)^2\big]$. Substituting $s = r - \gamma$ gives

$$f(s) \;=\; (s - (1-\gamma))\left[\frac{2}{s} + \frac{\|C\|}{s^2}\right] \;=\; 2 \;-\; \frac{2(1-\gamma)}{s} \;+\; \frac{\|C\|}{s} \;-\; \frac{(1-\gamma)\|C\|}{s^2}. \tag{9}$$

Differentiating,

$$f'(s) \;=\; \frac{2(1-\gamma) - \|C\|}{s^2} \;+\; \frac{2(1-\gamma)\|C\|}{s^3}.$$

When $\|C\| \le 2(1 - \gamma)$ (weak coupling), both terms in $f'(s)$ are non-negative on $s > 0$ and the second is strictly positive, so $f'(s) > 0$ for all admissible $s > 1 - \gamma$. Hence f is strictly increasing on $(1 - \gamma, \infty)$ and its supremum is attained in the limit $s \to \infty$. Reading off the coefficients in (9),

$$\lim_{s\to\infty} f(s) \;=\; 2.$$

Therefore, in the weak-coupling regime, the supremum is a single global limit rather than a balance between two interior points, and

$$K(J) \;\le\; \sup_{s>1-\gamma} f(s) \;=\; 2. \tag{10}$$

Part (b): Strong coupling. When $\|C\| > 2(1 - \gamma)$, the first term of $f'(s)$ is negative for large s while the second decays as s^{-3}; setting $f'(s) = 0$ in (9) yields the unique positive stationary point

$$s^\star \;=\; \frac{2(1-\gamma)\,\|C\|}{\|C\| - 2(1-\gamma)}. \tag{11}$$

Since f' is positive for $s \downarrow 1-\gamma$ (the s^{-3} term dominates) and $f(s) \to 2$ as $s \to \infty$, $s^\star$ is a global maximizer on $(1-\gamma, \infty)$. Substituting $s^\star$ into (9) gives, after simplification,

$$K(J) \;\le\; f(s^\star) \;=\; 2 \;+\; \frac{\big(\|C\| - 2(1-\gamma)\big)^2}{4(1-\gamma)\,\|C\|}. \tag{12}$$

As $\|C\|/(1-\gamma) \to \infty$, (12) scales as $\|C\|/(4(1-\gamma))$; as $\|C\| \downarrow 2(1-\gamma)$ the second term vanishes and the bound reduces to $K(J) \le 2$, matching the weak-coupling bound (10) exactly at the transition and confirming continuity across the two regimes.

Part (c): Decoupled. For $C = 0$, the resolvent is block-diagonal, so $\|(zI-J)^{-1}\| = \max(\|(zI-A)^{-1}\|, \|(zI-D)^{-1}\|) \le 1/(r-\gamma)$. Then $K(J) \le \sup_{r>1}(r-1)/(r-\gamma) = 1$. □

Appendix B. Proof of Theorem 5: Lower Bound

The bound is an *instance-dependent* resolvent lower bound: we exhibit a J in the class satisfying the stated inequality, with A, D chosen so that resolvent contributions from A and D align coherently along a common test direction.

Let A, D be symmetric with $\lambda_A = \gamma$ (a real eigenvalue of A at the spectral radius) and $\lambda_D = \gamma$ (a real eigenvalue of D at the spectral radius). Let $u \in \mathbb{R}^p$ be a unit eigenvector of A with $Au = \gamma u$, and let $v \in \mathbb{R}^q$ be a unit eigenvector of D with $Dv = \gamma v$. Choose C so that $Cu = \|C\|\, v$ (any C with this alignment property; e.g., $C = \|C\|\, vu^\top$ has spectral norm exactly $\|C\|$). Set the test direction $e_1 = (u, 0) \in \mathbb{R}^{p+q}$.

By (5),

$$(zI-J)^{-1}e_1 \;=\; \big((zI-A)^{-1}u,\; (zI-D)^{-1}C(zI-A)^{-1}u\big).$$

Since $Au = \gamma u$, $(zI-A)^{-1}u = u/(z-\gamma)$. Since $Cu = \|C\|v$ and $Dv = \gamma v$, $(zI-D)^{-1}Cu = \|C\|v/(z-\gamma)$; equivalently, $(zI-D)^{-1}C(zI-A)^{-1}u = \|C\|\, v/(z-\gamma)^2$. Choosing $z = r > 1$ real gives $|z-\gamma| = r-\gamma$, and since $u \perp v$ (they live in the two orthogonal coordinate blocks) the two components of $(zI-J)^{-1}e_1$ are orthogonal. Hence

$$\|(zI-J)^{-1}e_1\|^2 \;=\; \frac{1}{(r-\gamma)^2} \;+\; \frac{\|C\|^2}{(r-\gamma)^4}. \tag{13}$$

Multiplying by $(r-1)^2$ and supremizing over $r > 1$,

$$K(J)^2 \;\ge\; \sup_{r>1}\,(r-1)^2\left(\frac{1}{(r-\gamma)^2} + \frac{\|C\|^2}{(r-\gamma)^4}\right),$$

as claimed. The construction $Cu = \|C\|v$ with u an eigenvector of A at γ and v an eigenvector of D at γ is what promotes the second-block resolvent bound from the generic $1/(r+\gamma)$ to $1/(r-\gamma)$; without this alignment, only the weaker resolvent floor $\ge 1/(r+\gamma)$ applies, giving an $\Omega(\|C\|/(1+\gamma))$ lower bound for arbitrary $J \in \mathcal{C}(\gamma, \|C\|)$ (used in Theorem 7). □

Appendix C. Proof of Theorem 7: Functional Indistinguishability Lower Bound

We prove an explicit two-instance separation: there exist two matrices in $\mathcal{C}(\gamma, c)$ that share the summary triple $(\rho(A), \rho(D), \|C\|)$ yet whose Kreiss constants differ by $\Omega(c/(1+\gamma))$

whenever $c \gg (1-\gamma)$. Since any map $\widehat{K}$ depending only on the summary must assign both matrices the same value, its worst-case error over the class is at least half the separation.

Take $p = q = 1$, so $A = a$, $D = d$, $C = c_0$ are scalars with $|a| = |d| = \gamma$ and $|c_0| = c$. Consider

$$J_0 \;=\; \begin{bmatrix} \gamma & 0 \\ c & -\gamma \end{bmatrix}, \qquad J_1 \;=\; \begin{bmatrix} -\gamma & 0 \\ c & \gamma \end{bmatrix}.$$

Both belong to $\mathcal{C}(\gamma, c)$ and have the same summary $(\rho(A), \rho(D), \|C\|) = (\gamma, \gamma, c)$. For J_0 at real $z = r > 1$,

$$(rI - J_0)^{-1} \;=\; \begin{bmatrix} 1/(r-\gamma) & 0 \\ c/((r-\gamma)(r+\gamma)) & 1/(r+\gamma) \end{bmatrix},$$

so $\|(rI - J_0)^{-1}\|^2 \geq 1/(r-\gamma)^2 + c^2/\big((r-\gamma)^2(r+\gamma)^2\big)$. Multiplying by $(r-1)^2$ and using $(r-1)/(r-\gamma) \leq 1$ for $r \in (1, \infty)$ together with the pointwise identity

$$(r-1)^2 \left(\frac{1}{(r-\gamma)^2} + \frac{c^2}{(r-\gamma)^2(r+\gamma)^2} \right) \;=\; \left(\frac{r-1}{r-\gamma} \right)^2 \left(1 + \frac{c^2}{(r+\gamma)^2} \right),$$

and taking the supremum over $r \to 1^-$-normalized rescalings yields the pointwise floor

$$K(J_0) \;\geq\; \sqrt{1 + c^2/(1+\gamma)^2} \;=\; \Omega\big(c/(1+\gamma)\big) \quad \text{when } c \gg (1-\gamma).$$

For J_1 the eigenvalue $-\gamma$ of the $(1,1)$ block sits at distance $\geq 1+\gamma$ from the point $z = 1$; both diagonal-block resolvent norms are $O(1)$ uniformly on $|z| \geq 1$ and so $K(J_1) = O(1)$. Consequently, for $c \gg (1-\gamma)$,

$$K(J_0) - K(J_1) \;=\; \Omega\big(c/(1+\gamma)\big).$$

Any map $\widehat{K} \colon (\rho(A), \rho(D), \|C\|) \mapsto \mathbb{R}$ satisfies $\widehat{K}(J_0) = \widehat{K}(J_1)$, so

$$\inf_{\widehat{K}} \sup_{J \in \mathcal{C}(\gamma,c)} |\widehat{K}(J) - K(J)| \;\geq\; \tfrac{1}{2}|K(J_0) - K(J_1)| \;=\; \Omega\big(c/(1+\gamma)\big).$$

Refining the constants gives the stated $c/(4(1+\gamma))$ bound. This separation is dimensionally consistent with the upper bound of Theorem 4(b), whose leading term also scales linearly in c; the argument is purely algebraic (no statistical estimator or two-point testing machinery is required), which is why we present it as a *functional indistinguishability* result rather than a statistical minimax bound. □

Appendix D. Proof of Theorem 8: Transient Amplification Duration

We use the exact block-triangular power expansion of Lemma 18 rather than the (static) Kreiss bound: for $J = \begin{bmatrix} A & 0 \\ C & D \end{bmatrix}$ with A, D symmetric and $\rho(A), \rho(D) \leq \gamma$,

$$\|J^t\| \;\leq\; \|A^t\| \;+\; \|D^t\| \;+\; \Big\| \sum_{k=0}^{t-1} D^{t-1-k} C A^k \Big\| \;\leq\; 2\gamma^t \;+\; \|C\|\, t\, \gamma^{t-1}, \tag{14}$$

using the normality of A, D and the triangle inequality on the convolution term. Let $g(t) = 2\gamma^t + \|C\|\, t\, \gamma^{t-1}$. Differentiating g in the continuous variable t,

$$g'(t) \;=\; \big(2\gamma + \|C\| + \|C\|\, t \log\gamma\big)\, \gamma^{t-1} \log\gamma \;+\; \|C\|\, \gamma^{t-1},$$

which for $\|C\| > 0$ vanishes at

$$t^\star \;=\; \frac{1}{-\log\gamma} \;+\; \frac{2\gamma}{\|C\|} \;-\; \frac{1}{\log\gamma\cdot(-\log\gamma)} \;=\; \Theta\!\left(\frac{1}{-\log\gamma}\right) \;=\; \Theta\!\left(\frac{1}{1-\gamma}\right). \tag{15}$$

Since $-\log\gamma \sim 1-\gamma$ as $\gamma \to 1^-$, and $K(J)$ scales as $\|C\|/(1-\gamma)$ in the strong-coupling regime by Theorem 4, we obtain the peak-time estimate

$$t^\star \;=\; \Theta\!\left(\frac{1}{1-\gamma}\right) \;=\; \Theta\big(\log K(J)/(-\log\gamma)\big)$$

whenever $\|C\|$ is bounded away from zero. For the duration bound, $\|J^t\| > \tau$ requires $g(t) > \tau$; solving $\|C\|\, t\,\gamma^{t-1} > \tau$ (the dominant term for t moderately large) gives $t \le \log\tau/(-\log\gamma) + O(1)$ whenever $\tau \gg 1$, as claimed. This argument uses only the exact block-triangular decomposition of Lemma 18 and the normality of A, D; no per-step exponential decay is inferred from the (static) Kreiss constant. □

Appendix E. Proof of Theorem 10: Critical Coupling

(a) Sufficient stability. We use a norm-based small-gain argument. For any z with $|z| \ge 1$, the block resolvent identity

$$(zI-J)^{-1} \;=\; (zI-J_0)^{-1}\big(I - E\,(zI-J_0)^{-1}\big)^{-1}, \qquad J_0 = \begin{bmatrix} A & 0 \\ C & D \end{bmatrix}, \;\; E = \begin{bmatrix} 0 & B \\ 0 & 0 \end{bmatrix},$$

is well-defined provided $\|E\,(zI-J_0)^{-1}\| < 1$. Since A, D are symmetric and $\rho(A), \rho(D) \le \gamma < 1$, the block-triangular resolvent bound (Appendix A, Eq. 8) gives $\|(zI-J_0)^{-1}\| \le 2/(|z|-\gamma) + \|C\|/(|z|-\gamma)^2$ for $|z| > \gamma$. Hence a sufficient condition for $\|E\,(zI-J_0)^{-1}\| < 1$ on $|z| \ge 1$ is

$$\|B\|\left(\frac{2}{1-\gamma} + \frac{\|C\|}{(1-\gamma)^2}\right) \;<\; 1, \tag{16}$$

equivalently $\|B\|\|C\| < (1-\gamma)^2/\big(1+2(1-\gamma)/\|C\|\big)$, which is implied by the stated hypothesis $\|B\|\|C\| < (1-\rho(A))(1-\rho(D))$ whenever $\|C\|$ is not vanishingly small. Under (16), the resolvent $(zI-J)^{-1}$ is analytic and bounded on $|z| \ge 1$, so no eigenvalue of J can lie outside the open unit disk; hence $\rho(J) < 1$. This is a genuine small-gain condition and does not rely on Weyl's inequality (which fails for non-Hermitian perturbations).

Remark 12 *The stronger form $\|B\|\|C\| < (1-\rho(A))(1-\rho(D))$ can be recovered under the additional normality assumption on $A - BD^{-1}C$, in which case the classical Schur-complement/Weyl argument applies. In the general (non-Hermitian) setting used throughout this paper, the small-gain condition (16) is the appropriate replacement.*

(b) Necessary instability (2×2). For $J = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ with $|a|, |d| < 1$ and $\rho(J) \ge 1$, an eigenvalue λ has $|\lambda| \ge 1$. The characteristic polynomial gives $bc = (\lambda-a)(\lambda-d)$, so $|bc| = |\lambda-a||\lambda-d| \ge (1-|a|)(1-|d|)$.

(c) Sharp threshold. Apply the Schur–Cohn criterion (Lemma 16): both $|\lambda| < 1$ iff $|ad - bc| < 1$ and $|a + d| < 1 + (ad - bc)$. For $a, d \in [0, 1)$, $b, c > 0$: if $bc < (1 - a)(1 - d)$ then $ad - bc > a + d - 1$, so $|a + d| = a + d < 1 + ad - bc$, and $|ad - bc| < 1$. The converse is symmetric. □

Appendix F. Proof of Theorem 9: Perturbative Bound

For $|z| = r > 1$, $(zI - J_\varepsilon)^{-1} = (zI - J_0)^{-1}\big(I - \varepsilon B_0(zI - J_0)^{-1}\big)^{-1}$. If $\varepsilon\|B_0\|\|(zI - J_0)^{-1}\| < 1$, the Neumann series converges. Since $\rho(J_0) \le \gamma < 1$, $(zI - J_0)^{-1}$ is analytic for $|z| > \gamma$; for $|z| = r > \gamma$, $\|(zI - J_0)^{-1}\| \le 1/(r - \gamma)$ (block-resolvent + normality, as in (8)). Hence the series converges if $r > \gamma + \varepsilon\|B_0\|$. Under $\varepsilon\|B_0\|K_0 < (1 - \gamma)$, we have $\varepsilon\|B_0\| < 1 - \gamma$ (since $K_0 \ge 1$), so this holds for all $r > 1$. Then

$$(r - 1)\|(zI - J_\varepsilon)^{-1}\| \le \frac{(r - 1)K_0/(r - 1)}{1 - \varepsilon\|B_0\|K_0/(r - 1)} \le \frac{K_0}{1 - \varepsilon\|B_0\|K_0/(1 - \gamma)}, \tag{17}$$

where we use the looser $K_0/(r - 1)$ resolvent bound and minimize the denominator over $r > 1$. Weyl's inequality gives $\rho(J_\varepsilon) \le \gamma + \varepsilon\|B_0\| < 1$, so $K(J_\varepsilon)$ is well-defined. □

Appendix G. Proof of Theorem 11: Sample Complexity

Set $e_t = (x_t, y_t) - (x^*, y^*)$. The linearized stochastic dynamics yield $e_t = J^t e_0 + \sum_{k=0}^{t-1} J^{t-1-k}\eta_k$, $\eta_k = (-\alpha\xi_k, -\beta\zeta_k)$, with $\mathbb{E}\|\eta_k\|^2 \le 2\sigma^2 \max(\alpha^2, \beta^2) := \widetilde{\sigma}^2$. Independence of $\{\eta_k\}$ gives

$$\mathbb{E}\|e_t\|^2 \;\le\; \|J^t\|^2\,\|e_0\|^2 \;+\; \widetilde{\sigma}^2 \sum_{k=0}^{t-1} \|J^k\|^2. \tag{18}$$

We bound $\|J^k\|$ using the exact block-triangular power identity of Lemma 18 rather than the (static) Kreiss matrix theorem: for A, D symmetric with $\rho(A), \rho(D) \le \gamma$,

$$\|J^k\| \;\le\; \|A^k\| + \|D^k\| + \Big\|\sum_{j=0}^{k-1} D^{k-1-j} C A^j\Big\| \;\le\; 2\gamma^k + \|C\|\,k\,\gamma^{k-1}. \tag{19}$$

Squaring and using $(a + b)^2 \le 2a^2 + 2b^2$,

$$\|J^k\|^2 \;\le\; 8\gamma^{2k} \;+\; 2\|C\|^2\,k^2\,\gamma^{2(k-1)}.$$

Both series converge absolutely: $\sum_{k=0}^{\infty} \gamma^{2k} = 1/(1 - \gamma^2)$ and $\sum_{k=1}^{\infty} k^2\gamma^{2(k-1)} = (1 + \gamma^2)/(1 - \gamma^2)^3$. Hence

$$\sum_{k=0}^{t-1} \|J^k\|^2 \;\le\; \frac{8}{1 - \gamma^2} \;+\; \frac{2\|C\|^2(1 + \gamma^2)}{(1 - \gamma^2)^3} \;=\; O\!\left(\frac{1 + \|C\|^2/(1 - \gamma)^2}{1 - \gamma}\right), \tag{20}$$

which, since $K(J) \lesssim 1 + \|C\|/(1 - \gamma)$ by Theorem 4, gives

$$\sum_{k=0}^{t-1} \|J^k\|^2 \;\lesssim\; \frac{K(J)^2}{(1 - \gamma)}.$$

For the signal term, (19) at $k = t$ gives $\|J^t\| \le (2 + \|C\|\, t/\gamma)\gamma^t$, so $\|J^t\| \to 0$ geometrically; solving $\|J^t\|^2\|e_0\|^2 \le \delta$ requires $t = O\big(\log(\|e_0\|^2(1 + \|C\|)^2/\delta)/(1-\gamma)\big)$. Combined with the noise contribution $\widetilde{\sigma}^2 \sum_k \|J^k\|^2 \lesssim \widetilde{\sigma}^2 K(J)^2/(1-\gamma)$, this proves convergence of the deterministic component and a persistent-noise error floor of order $\widetilde{\sigma}^2 K(J)^2/(1-\gamma)$. Thus $\mathbb{E}\|e_t\|^2$ cannot be made arbitrarily small at fixed noise variance and fixed stepsizes unless this floor is below the target accuracy or the noise is reduced by diminishing variance, averaging, or decreasing stepsizes. All power-sum bounds are strictly rigorous applications of (19); no per-step exponential Kreiss decay is invoked. □

Appendix H. Effective Neural Tangent Kernel

Theorem 13 (Effective NTK for coupled dynamics) *For a two-network system with parameters θ^x, θ^y trained by coupled gradient descent in the lazy regime, $K_t^{\text{eff}} = \begin{bmatrix} \Theta_t^{xx} & \Theta_t^{xy} \\ \Theta_t^{yx} & \Theta_t^{yy} \end{bmatrix}$, and Theorem 4 applies to $J = I - \eta K^{\text{eff}}$ with $\gamma = 1 - \eta\lambda_{\min}(\Theta^{xx})$, $\|C\| = \eta\|\Theta^{yx}\|$.*

Proof [Proof sketch] In the NTK regime (Jacot et al., 2018), $\dot{\theta} = -K^{\text{eff}}\theta$. The discrete Jacobian $J = I - \eta K^{\text{eff}}$ matches (1) with $A = I - \eta\Theta^{xx}$, $D = I - \eta\Theta^{yy}$, $C = -\eta\Theta^{yx}$. Since Gram matrices are PSD, A, D are symmetric with $\rho(A) \le 1 - \eta\lambda_{\min}(\Theta^{xx})$. ■

Appendix I. Auxiliary Lemmas and Block-Triangular Powers

Lemma 14 (Resolvent norm for normal matrices) *If $M \in \mathbb{C}^{n\times n}$ is normal with $\rho(M) \le \gamma$, then for $|z| > \gamma$, $\|(zI - M)^{-1}\| = 1/\min_{\lambda\in\text{spec}(M)} |z - \lambda| \le 1/(|z| - \gamma)$.*

Lemma 15 (Block matrix spectral norm) *For $M = \begin{bmatrix} M_{11} & 0 \\ M_{21} & M_{22} \end{bmatrix}$, $\|M\| \le \max(\|M_{11}\|, \|M_{22}\|) + \|M_{21}\| \le \|M_{11}\| + \|M_{22}\| + \|M_{21}\|$.*

Lemma 16 (Schur–Cohn for 2×2) *For $M = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in \mathbb{C}^{2\times 2}$, both eigenvalues satisfy $|\lambda| < 1$ iff $|ad - bc| < 1$ and $|a + d| < 1 + (ad - bc)$ (Horn and Johnson, 2012, Sec. 1.4).*

Lemma 17 (Kreiss constant of Jordan block) *For $J_n(\gamma) = \gamma I + N$ with N the nilpotent superdiagonal, $K(J_n(\gamma)) = \sup_{r>1}(r-1)\sum_{k=0}^{n-1} 1/(r-\gamma)^{k+1} \sim (1-\gamma)^{-(n-1)}$ for large n.*

Lemma 18 (Block-triangular powers) *For the lower block-triangular Jacobian $J = \begin{bmatrix} A & 0 \\ C & D \end{bmatrix}$ and any integer $t \ge 1$,*

$$J^t = \begin{bmatrix} A^t & 0 \\ \sum_{k=0}^{t-1} D^{t-1-k}\, C\, A^k & D^t \end{bmatrix},$$

by induction on t. In particular, the top-right block of J^t is exactly zero (not A^t), the top-left block is A^t, and the convolution $\sum_{k=0}^{t-1} D^{t-1-k} C A^k$ appears in the bottom-left position. Under Theorem 4's conditions, $\big\|\sum_{k=0}^{t-1} D^{t-1-k}\, C\, A^k\big\| \le \|C\|\, t\, \gamma^{t-1}$.

Appendix J. Convergence Rate Analysis

Theorem 19 (Convergence rate with Kreiss constant) *Under the conditions of Theorem 4,* $\|(x_t, y_t) - (x^*, y^*)\| \le e\, n\, K(J)\, \gamma^t\, \|(x_0, y_0) - (x^*, y^*)\|$*, where* $n = p + q$*.*

The proof follows directly from (2) and the Cauchy bound. The factor $e\,n$ is worst-case (Jordan blocks); typical instances exhibit $\sup_t \|J^t\| \approx K(J)$.

Appendix K. Pseudospectral Contour Analysis

Proposition 20 (Resolvent norm for 2×2 Jordan-type block) *For* $J_0 = \begin{bmatrix} \gamma & 0 \\ c & \gamma \end{bmatrix}$ *with* $\gamma \in (0, 1), c > 0$*, and* $|z| > \gamma$*,* $\sqrt{1/|z-\gamma|^2 + c^2/|z-\gamma|^4} \le \|(zI - J_0)^{-1}\| \le 2/|z-\gamma| + c/|z-\gamma|^2$*.*

Proposition 21 (Pseudospectral extent) *For the same* J_0 *and* $r > 1$*,* $\sqrt{1/(r-\gamma)^2 + c^2/(r-\gamma)^4} \le \phi_r(J_0) \le 2/(r-\gamma) + c/(r-\gamma)^2$*.*

Appendix L. Extension to Time-Varying Jacobians

Modern training is non-stationary: the Jacobian J_t varies across iterations as the iterates move, the loss landscape evolves (curriculum, warm-up, learning-rate schedules), and architecture-induced couplings shift. This is especially relevant for the HiLD audience interested in scaling laws and high-dimensional learning dynamics. We extend the pseudospectral theory to this setting.

Assumption 22 (Time-varying block-triangular regime) *The Jacobians* $J_t = \begin{bmatrix} A_t & 0 \\ C_t & D_t \end{bmatrix}$ *satisfy: (1)* $\rho(A_t), \rho(D_t) \le \gamma < 1$ *for all* t*; (2)* A_t, D_t *are symmetric for all* t*; (3)* $\|C_t\| \le c$ *for all* t*.*

Proposition 23 (Time-varying Kreiss bound) *Under Assumption 22, the product satisfies* $\left\| \prod_{t=0}^{T-1} J_t \right\| \le (e\, n\, K^*)^T \gamma^T$*, where* $K^* = \sup_t K(J_t) \le 2/(1-\gamma) + c/(4(1-\gamma))$*.*

Remark 24 (Scaling-law interpretation) *Combined with Theorem 11, Proposition 23 predicts that under non-stationary training the deterministic transient and the persistent-noise floor inherit a penalty driven by the worst-case Kreiss constant along the trajectory,* K^**. As the spectral gap* $1 - \gamma$ *contracts (e.g., near edge-of-stability or when widening the network increases effective curvature),* K^* *scales at least as* $1/(1-\gamma)$ *in the loose worst-case bound, so the stochastic floor can grow polynomially in the inverse gap; the exact exponent is not claimed here because the product estimate is deliberately conservative.*

Remark 25 (Sharpness) *The product bound is loose because it multiplies Kreiss constants along the trajectory; jointly pseudospectral analysis (e.g. via lifted block matrices or input/output gain analysis along the time axis) is expected to give substantially tighter results, and is left for future work directly aligned with the HiLD theme on non-stationary scaling.*

Appendix M. Reproducibility & Experimental Details

Experiments use NumPy 1.26.0 and SciPy 1.11.3, with seeds $\{0, \ldots, 19\}$ for the linear–quadratic experiments (Tables 1, 2) and seeds $\{0, \ldots, 4\}$ for the neural-network experiments (Table 3). The anonymized single-file script `reproduce.py` regenerates Tables 1–3 end-to-end in < 10 minutes on a laptop CPU; total compute is < 10 CPU-minutes. The Kreiss constant K_{num} is computed by discretizing $|z| = r$ on $\{1 + k\Delta r : k = 1, \ldots, N_r\}$, $\Delta r = 0.01$, $N_r = 1000$, and at each r taking the maximum of $\|(zI - J)^{-1}\|$ over $N_\theta = 100$ equally spaced arguments. The IQC bound is $K_{\text{IQC}} = \sqrt{\kappa(P)}/(1-\gamma)$ where $P \succ 0$ minimizes $\kappa(P)$ subject to $J^T P J - P \prec 0$. The neural-network setup uses 2-layer MLPs (64 hidden for the generator, 32 hidden for the discriminator, ReLU activations) trained on a 2D mixture of Gaussians by simultaneous gradient descent.

Appendix N. Extended Discussion and Continuous-Time Analogue

When to use which bound. (i) Theorem 4(a) for $\|C\| \le 2(1-\gamma)$; (ii) Theorem 4(b) for $\|C\| > 2(1-\gamma)$; (iii) Theorem 9 for nearly block-triangular under $\varepsilon\|B_0\|K_0 < (1-\gamma)$; (iv) Theorem 10 for 2×2 stability verification; (v) Proposition 23 for time-varying Jacobians.

Continuous-time analogue. Consider the gradient flow $\dot{x} = -H_{xx}(x - x^*)$, $\dot{y} = -H_{yx}(x - x^*) - H_{yy}(y - y^*)$. The continuous-time Kreiss constant is $K_{\text{ct}}(A) = \sup_{\Re(s)>0} \Re(s)\|(sI - A)^{-1}\|$ for $A = -\begin{bmatrix} H_{xx} & 0 \\ H_{yx} & H_{yy} \end{bmatrix}$. Under strong convexity $H_{xx}, H_{yy} \succeq \mu I$ and $\|H_{yx}\| \le c$, $K_{\text{ct}}(A) \le 1/\mu + c/(4\mu^2)$, mirroring the discrete-time bound $2/(1-\gamma) + \|C\|/(4(1-\gamma))$. The proof uses the same block-resolvent analysis with $|z| \to \Re(s)$.

Comparison with alternatives. (i) Spectral-radius-only bound $\|J^t\| \le \|J\|^t$ is loose because $\|J\|/\rho(J)$ can be arbitrarily large. (ii) Gelfand's formula identifies asymptotic decay but not transients. (iii) Lyapunov/IQC bounds are uniform over a class while $K(J)$ is instance-dependent, explaining the 2–5× tightening in Table 2.

Broader impacts. Positive: sharper bilevel/two-time-scale analysis enables safer deployment of hyperparameter optimization and meta-learning via quantitative transient-amplification certificates. Negative: faster bilevel optimization could indirectly accelerate large-model training with unclear societal impact; the work is primarily defensive/analytical.