When Averaging Amplifies: Stability and Sensitivity in Adaptive Mixing

Anonymous authors
Paper under double-blind review

Abstract

Positive averaging can converge while amplifying input changes. We study adaptive mixing that retains the original input, as in non-local-means RED denoising and related graph and equilibrium architectures. A class-sharp convergence certificate gives an optimal input-response bound: smooth stochastic mixers approach that bound while their frozen averages (realized weights held fixed) approach identity, and cross-token responses can be negative. A photographic RED/NLM example and ordinary softmax kernels exhibit amplification. Convergence also need not preserve predictions. At a prespecified strong-mixing setting, adaptive mixing of full-dimensional DINO features in 32-image contexts reduces accuracy on 512 matched CIFAR-10 queries from 88.09% to 35.87%; freezing input weights retains 88.02%. An exact score decomposition and context removal identify recoverable interference: the original classifier recovers 87.96%. Finally, intermediate iteration budgets can destabilize downstream feedback even when a longer solve is stable. We give a sufficient condition covering every budget. These results separate guarantees for convergence, input response, task fidelity, and downstream use.

1 Introduction

Positive averaging keeps values within their convex hull. Many iterative systems, however, rebuild the weights from the values being mixed while retaining an original input. Changes to that input alter both values and future weights. A settled computation can therefore be much more sensitive than its realized averaging matrix suggests.

In regularization by denoising (RED), a non-local-means filter rebuilds image-patch similarities from the evolving estimate while each fixed-point update retains the observed image (Romano et al., 2017; Reehorst & Schniter, 2019). Graph networks based on the p-Laplacian also rebuild edge weights and reinject initial features, with node-dependent retention (Fu et al., 2022). Nonlinear GRAND++ supplies a fixed input-derived source to adaptive diffusion; equilibrium Transformers repeatedly combine current-state attention with injected input (Thorpe et al., 2022; Bai et al., 2019). We study their shared feedback mechanism through a positive-mixing update, exact for RED/NLM denoising and requiring additional terms or bounds for the other architectures.

Our central result separates *the weights that represent an output* from *the derivative that describes its response*. Every equilibrium is a positive average of the original inputs. Nevertheless, smooth contractive mixers can have negative cross-token derivatives and attain the class-optimal response bound while their frozen averages approach identity. Figure 1 exhibits amplification and distinct attracting states using a published denoising rule on photographic data. The sufficient contraction argument is classical; its sharpness within the stochastic-builder class and the extremal positive-mixer construction are the specific theoretical contributions.

The distinction extends to downstream use in two ways. First, a fixed classifier can lose class margins after converged mixing. In a matched-query image experiment, we trace a 52.21 percentage-point loss to recoverable context interference, using frozen weights, changed companions, and context removal as controls. Feature-nearest companions still leave an 8.59-point loss (Figure 3). Second, a downstream feedback loop receives a finite computation. Its sensitivity can peak at intermediate iteration budgets, making the outer loop unstable even when a longer computation is stable. The same contraction margin supplies one sufficient condition for every budget.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

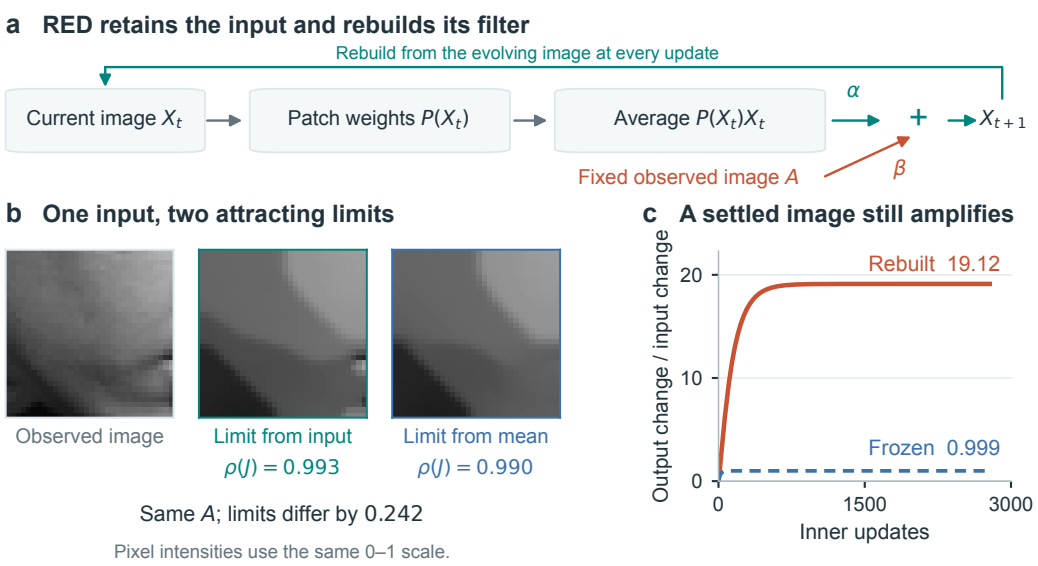


Figure 1: **A converged denoising loop can amplify its input.** (a) RED/NLM rebuilds patch similarities and retains the observed image (Romano et al., 2017). (b) The same photographic input gives distinct attracting limits from input and mean starts. (c) A perturbation of at most 10^{-5} per pixel produces 19.12-fold amplification, versus 0.999 with equilibrium weights frozen. Its direction maximizes the local maximum-row gain (19.23); it is a selected probe, not sampled noise. Here $\alpha = 0.95$, bandwidth 0.025. Appendix A.7.2 gives the full parameter screen, derivative checks, and local scope.

Relation to prior work. Adaptive filter weights and the filter Jacobian are already distinguished in RED analysis (Reehorst & Schniter, 2019); input-injected attention admits classical fixed-point guarantees (Vuckovic et al., 2021). Our contribution is the class-sharp convergence boundary, an optimal response construction approaching identity in its frozen values, and controlled consequences for class margins and outer feedback. Appendix A.6 relates these results to existing guarantees; Appendix A.7 distinguishes exact instances from architectural extensions.

2 A SHARP CONVERGENCE CERTIFICATE

Let $A \in \mathbb{R}^{N\times d}$ contain the fixed input representations, called *anchors*, and X_t the evolving states. A *builder* $\mathcal{P}(X)$ returns nonnegative weights whose rows sum to one. We study

$$X_{t+1} = F_A(X_t) = \alpha\mathcal{P}(X_t)X_t + \beta A, \qquad \beta = 1-\alpha, \quad 0 \le \alpha < 1. \tag{1}$$

In RED denoising, A is the observed image, X_t its current estimate, and $\mathcal{P}(X_t)$ the row-normalized non-local-means weights. The published fixed-point rule with identity observation operator is exactly (1), with $\alpha = \lambda\sigma^2/(1+\lambda\sigma^2)$ for regularization weight λ and observation-noise variance σ^2 (Romano et al., 2017). Fixed-graph propagation such as APPNP is the constant-builder case (Gasteiger et al., 2019). The changing weights, rather than input retention alone, supply the additional feedback.

Use the maximum-row norm $\|X\|_{\infty,2} = \max_i \|x_i\|_2$. Fix $R \ge \|A\|_{\infty,2}$, $R > 0$, and let $B_R = \{X : \|X\|_{\infty,2} \le R\}$. Suppose

$$\|\mathcal{P}(X) - \mathcal{P}(Y)\|_\infty \le L_P \|X - Y\|_{\infty,2}, \qquad \kappa_f = RL_P, \qquad m = \beta - \alpha\kappa_f, \tag{2}$$

for $X, Y \in B_R$, where the matrix norm is the maximum absolute row sum. The dimensionless quantity κ_f measures how strongly changed weights can act on bounded representations; m is the certified contraction margin. Any state-dependent mask, normalization, or bandwidth belongs to the builder and its bound.

Theorem 1 (Convergence and class sharpness)**.** *The ball B_R is invariant under* (1)*. If $m > 0$, the update is a contraction with upper factor $q = \alpha(1 + \kappa_{\mathrm{f}}) = 1 - m < 1$, and every initialization in B_R converges geometrically to its unique fixed point. For every $\alpha \in (0, 1)$ and $\varepsilon > 0$, there are a smooth row-stochastic builder with $N = 2$, $d = 1$, $R = 1$, $\beta < \alpha L_P \leq (1 + \varepsilon)\beta$, and an anchor with $\|A\|_{\infty,2} = 1$ such that* (1) *has at least two locally attracting fixed points in B_1.*

Values and weights contribute separately. Split the difference of two updates into $\alpha\mathcal{P}(X)(X - Y)$ and $\alpha(\mathcal{P}(X) - \mathcal{P}(Y))Y$. Their norms are bounded by $\alpha \|X - Y\|_{\infty,2}$ and $\alpha R L_P \|X - Y\|_{\infty,2}$, respectively. Banach's theorem then gives convergence.

Why the threshold is sharp. Two scalar states suffice. A smooth mixer can make their contrast map cross the identity three times, with two attracting roots and a repelling root between them, while αL_P approaches β from above. Their mean contracts independently (Appendix A.2.2). This is sharpness over the builder class, not a claim that every mixer beyond the boundary fails. Figure 1 illustrates distinct attracting states in a Gaussian NLM kernel; it is not a kernel-specific sharpness result.

On a common invariant ball, two anchors satisfy

$$\|X^*(A') - X^*(A)\|_{\infty,2} \leq \frac{\beta}{m} \|A' - A\|_{\infty,2} . \tag{3}$$

A candidate's solution error is at most its update residual divided by m. Both conclusions require a valid domain bound, not a sampled derivative or an observed convergence rate.

Applying the certificate. For dot-product softmax, $\mathcal{P}(X)_{ij} = \mathrm{softmax}_j(s\langle x_i, x_j\rangle)$, we have $L_P \leq 2sR$. Thus $2\alpha s R^2 < 1 - \alpha$ certifies convergence. Stronger anchoring, softer logits, or a tighter invariant domain improve the test. With input initialization, all rows remain in the anchor convex hull C. Since builder differences have zero row sums, RL_P can be replaced by $D_C L_P/2$, where D_C is the hull diameter (Appendix A.2). Perturbed anchors still require a common domain.

The unit ball certifies the dot-product setting $(\alpha, s) = (0.6, 0.3)$ with $m = 0.04$. It does not certify the strong-mixing image failure or the ordinary-kernel response maximum below. Context-specific hulls certify 1,349 of 4,608 nonidentity image confirmation configurations, all dot-product (Appendix A.4.5). We distinguish certified convergence from numerical settling throughout.

3 Positive values, signed and amplified response

Operator norms below are induced by the maximum-row norm unless stated otherwise. At an equilibrium, write $P_* = \mathcal{P}(X^*)$. Holding these realized weights fixed gives

$$X^* = S_* A, \qquad S_* := \beta(I - \alpha P_*)^{-1} = \sum_{k=0}^{\infty} \beta \alpha^k P_*^k. \tag{4}$$

The coefficients in this series are nonnegative and sum to one. Consequently S_* is row stochastic and nonexpansive in the maximum-row norm: each output is a convex combination of the anchors.

The response to changing the anchors includes the adaptive-filter product rule emphasized by Reehorst & Schniter (2019). On a differentiable equilibrium branch with invertible $I - J$,

$$\begin{aligned} J[U] &= \underbrace{\alpha P_* U}_{\text{changing values}} + \underbrace{\alpha(D\mathcal{P}(X^*)[U])X^*}_{\text{changing weights}}, \\ H &:= D_A X^* = \beta(I - J)^{-1}. \end{aligned} \tag{5}$$

Local attraction, $\rho(J) < 1$, suffices. Unlike S_*, H need not be nonnegative or nonexpansive. Contraction compares trajectories with the *same* anchors; input response compares equilibria with *different* anchors. Rebuilding can therefore make an increase in one input decrease another output without violating positivity of any averaging weight.

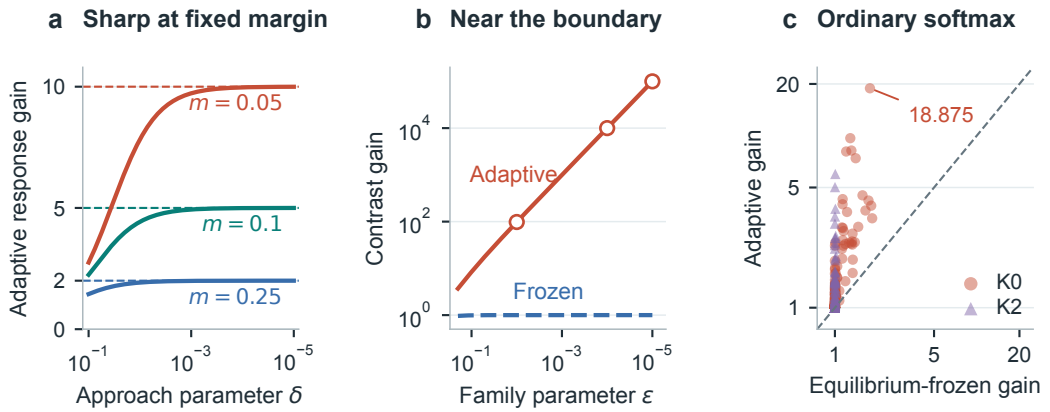


Figure 2: **The sharp response bound is compatible with almost unchanged frozen values.** (a) Smooth two-token families approach β/m at fixed contraction margins; dashed lines are the respective bounds, with $\alpha = \beta = 1/2$. (b) A separate family approaches the convergence boundary: adaptive response diverges while frozen contrast retention approaches one. Open circles are independent 80-digit finite differences. (c) All 360 small-cloud confirmation configurations for dot-product (K0) and adaptive-bandwidth Gaussian (K2) softmax. The maximum Euclidean response is 18.875, versus 1.766 with its equilibrium weights frozen. The diagonal denotes equal gain. These designed configurations establish possibility, not prevalence.

Proposition 1 (Optimal response from positive mixers)**.** *For every $0 < \alpha < 1$ and $0 < m < \beta$, there are smooth two-token builders contractive on a common unit ball with upper factor $1 - m$, and interior anchors, such that $S_* \to I$ and $\|D_A X^*\| \to \beta/m$ (in both induced maximum-row and Euclidean norms). Cross-token derivatives are negative sufficiently far along the family. With $\alpha = \beta = 1/2$, a further family has $m \downarrow 0$ and $\|D_A X^*\| \to \infty$ while $S_* \to I$.*

Thus the bound in (3) is optimal even at a fixed positive margin. For example, $\alpha = 1/2$ and $m = 1/4$ permit response approaching two with contraction factor at most $3/4$ and frozen smoother approaching identity. Unbounded response requires a vanishing margin. In the explicit diverging family at $\varepsilon = 0.01$, frozen contrast retention is 0.999900, response gain is 98.049, and the cross-token derivative is -48.524. Their linear-response neighborhood shrinks as the boundary is approached (Appendix A.2.5).

How a positive mixer attains the bound. The construction separates the mean of two tokens from their contrast. Let $E = \mathbf{1}\mathbf{1}^\top/2$ project onto their common value and $Q = I - E$ onto their difference. For anchors $A = a(1, -1)^\top$, $0 < a < 1$, choose $0 < \mu < 1$ and set

$$k = \frac{\beta - m}{\alpha}, \qquad z_* = \frac{\beta a}{1 - \alpha\mu}, \qquad \mathcal{P}(X) = E + g\left(\frac{x_1 - x_2}{2}\right) Q, \tag{6}$$

where $g(z)$ is a smooth clipping of $\mu + k(z - z_*)$ to $[0, 1]$, equal to that affine function near z_*. Every entry of $\mathcal{P}$ is $(1 + g)/2$ or $(1 - g)/2$, so the builder is nonnegative and stochastic, with $L_P = k$. The unit ball is a common contraction domain, including a neighborhood of A. At the equilibrium $X^* = z_*(1, -1)^\top$,

$$S_* = E + \frac{\beta}{1 - \alpha\mu} Q, \qquad H = E + \underbrace{\frac{\beta}{1 - \alpha\mu - \alpha k z_*}}_{G} Q. \tag{7}$$

The weight derivative contributes the extra term $\alpha k z_*$ in the response denominator. As $a, \mu \uparrow 1$, the frozen contrast coefficient tends to one while $G \to \beta/m$. Each off-diagonal entry of H is $(1 - G)/2$, which becomes negative. Changing one anchor can redistribute positive weights enough to outweigh its direct contribution; Appendix A.1 gives a numerical example on a certified common domain. Appendix A.2.5 specifies the smoothing and proves both limits.

Ordinary softmax also amplifies. We evaluate tied dot-product weights (K0) and Gaussian weights whose bandwidth depends on the evolving cloud (K2). Both apply rowwise softmax; K2 uses logits

$$\ell_{ij}^{(2)} = -\frac{s\,\|x_i - x_j\|_2^2}{2h^2(X)}, \qquad h^2(X) = 0.01 + \frac{\sum_{i\neq j}\|x_i - x_j\|_2^2}{N(N-1)}. \tag{8}$$

We differentiate the complete builder, including its bandwidth. All geometry trajectories settle from both prescribed starts; 92 of 900 confirmation configurations reach distinct endpoints. The largest dense response is 18.875 despite $\rho(J) = 0.9397$: a bridge cloud with K0, $N = 16, d = 8, \alpha = 0.6, s = 30$ in the confirmation grid. Freezing P_* gives 1.766; rebuilding once from each perturbed input and then holding the weights fixed gives 3.721. An independent central difference agrees with the adaptive derivative to relative error 3.97×10^{-6}.

These are Euclidean operator norms; a row-stochastic matrix can exceed one in that norm. Along the same selected direction, the maximum-row gain is 15.559 adaptive versus 0.821 frozen, so the separation also holds in the theorem's norm. Re-solved finite perturbations remain amplified (Appendix A.3.2). The example is an exploratory maximum. The full response inverse, not the scalar proxy $\beta/(1-\rho(J)) = 6.635$, determines its sensitivity.

Amplification at finite perturbation sizes. For the same response maximum, let V be the unit Frobenius-norm direction of greatest local response. Independently solving at $A \pm hV$ gives central gain $\|X^*(A+hV) - X^*(A-hV)\|_F/(2h)$ of 18.882 at $h = 10^{-4}$, close to the derivative's 18.875. It reaches 45.946 near $h = 1.78 \times 10^{-3}$ and remains 3.930 at $h = 0.1$, versus a frozen directional gain of 0.869 throughout. Starting each perturbed solve from either its anchor or the nominal equilibrium gives gains agreeing within 2×10^{-9}. Thus the amplification survives finite changes, although its size is nonlinear and the local derivative does not describe the entire excursion (Appendix A.3.2).

4 Converged mixing can overturn class margins

Input sensitivity and classification ask different questions. Sensitivity compares nearby inputs; a class margin compares competing scores at one input. Cross-image aggregation can improve transductive classification, yet repeated feature re-representation reduces accuracy relative to one pass (Cui & Guo, 2021). Graph learning under heterophily likewise benefits from separating own and neighbor representations (Zhu et al., 2020). We isolate the role of adaptive rebuilding using fixed DINO ViT-S/16 features (Caron et al., 2021), fixed affine heads, and CIFAR-10 (Krizhevsky, 2009). The experiments mix *images within a context*; the encoder and classifier stay unchanged.

A matched-query intervention. From a documented extraction of all official images, we prepare full 384D features and a 16D PCA branch. Training-only centering, scaling, and the invertible map $z \mapsto z/\sqrt{1+\|z\|_2^2}$ bound the states. Each branch fits its own ridge head. Each of 512 matched test queries receives 31 companions: three uniform draws or one deterministic nearest-neighbor context. Both branches retrieve in the training-fitted 16D features, without labels. Only queries are scored; all methods share companions. Uniform results average draws within query.

The focal parameters, $\alpha = 0.95, s = 30$, were selected in the preceding image campaign and fixed before this study. Adaptive accuracy falls from 88.09% to 35.87% with uniform companions, a loss of 52.21 percentage points[1] (pp; paired query-bootstrap 95% interval $[48.96, 55.53]$). Building $P(A)$ once and freezing it retains 88.02%. Each uniform draw exhibits the loss: adaptive accuracy ranges from 33.98% to 37.30%. Feature neighbors raise adaptive accuracy to 79.49%, but still leave an 8.59-point loss ($[5.27, 11.91]$); frozen weights retain 87.70%. Same-label fractions are 9.88% for uniform and 65.93% for nearest companions. Feature-based selection reduces interference without eliminating it. The intervals condition on this feature bank and context design.

[1] Accuracy differences are computed from exact counts before rounding.

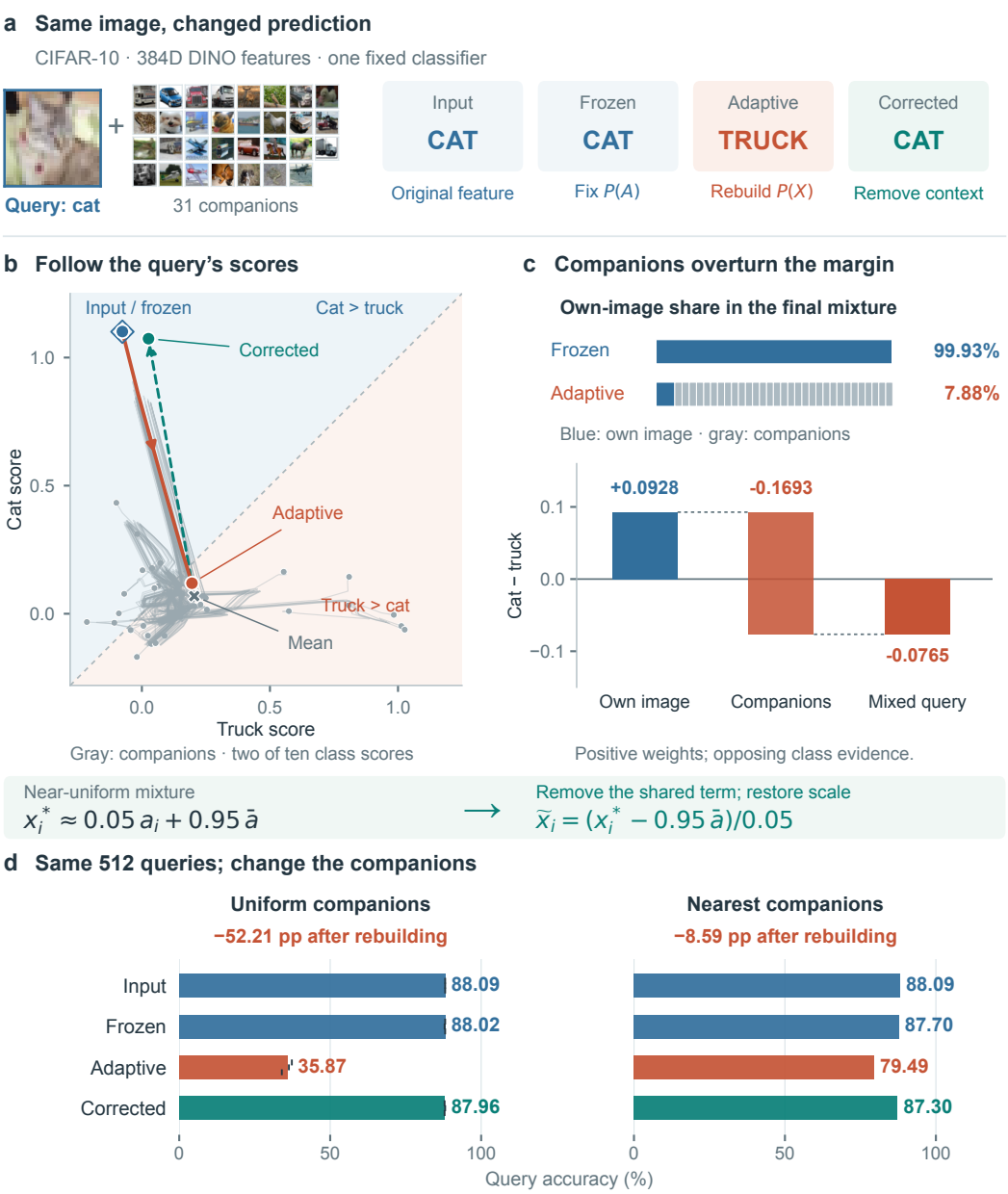


Figure 3: **Adaptive mixing hides class evidence that context removal recovers.** Full 384D DINO features, $\alpha = 0.95, s = 30$. (a) First stored query and 31 uniform companions; mixing acts on features. One ten-class head predicts cat with input or frozen weights, truck after rebuilding, and cat after correction. (b) Actual trajectories in two class scores; dashed arrow: correction, cross: original context mean. (c) Positive equilibrium-smoother weights carry opposing class evidence: companions outweigh the query's own cat-minus-truck contribution (9). (d) All 512 matched queries. Uniform accuracy averages three draws within query (marks show each draw); nearest companions use feature similarity without labels. Both losses are relative to input. The correction uses the mean $\bar{a}$ of all 32 original features (11): exact under uniform weights, diagnostic at the measured, approximately uniform equilibria. All focal solves settle.

No dimensional truncation is needed for this effect. The 16D branch also loses 46.48 pp. Table 1 places the full-dimensional result in the complete parameter comparison. At $\alpha = 0.85$, rebuilding already reduces uniform-context accuracy to 76.11%, while freezing retains 88.09%. Softer logits and stronger anchoring produce much smaller changes.

Table 1: **The full-dimensional loss grows under stronger adaptive mixing.** Accuracy (%) on the same 512 queries; input accuracy is 88.09%. Uniform results average three draws within query. Frozen weights are built at the input. The first two settings are certified on the unit ball. The dagger marks one capped, unresolved solve retained in accuracy; all other adaptive solves settle. Correction at $(0.95, 30)$ reaches 87.96% with uniform and 87.30% with nearest companions.

		Uniform companions		Nearest companions	
α	s	Frozen	Adaptive	Frozen	Adaptive
0.30	0.3	88.28	88.28	87.89	87.89
0.60	0.3	87.37	87.30	84.96	84.96
0.85	30	88.09	76.11	88.48	80.27†
0.95	30	88.02	35.87	87.70	79.49

The certified setting $(0.6, 0.3)$ loses 0.78 pp with uniform companions and 3.13 pp with neighbors. Certification ensures convergence and uniqueness, not class-margin preservation. Appendix A.4 gives both dimensions, all context draws, and paired intervals.

Exact accounting of whose scores survive. For an affine classifier, let $Z_0 = AW + \mathbf{1}b^\top$. Since $S_*\mathbf{1} = \mathbf{1}$, the output scores satisfy $Z^* = S_* Z_0$. A true-class-versus-competitor margin is therefore

$$Z^*_{iy} - Z^*_{ik} = \underbrace{(S_*)_{ii}(Z_{0,iy} - Z_{0,ik})}_{\text{own-image contribution}} + \underbrace{\sum_{j \neq i} (S_*)_{ij}(Z_{0,jy} - Z_{0,jk})}_{\text{companion contribution}}. \tag{9}$$

In Figure 3, rebuilding reduces the query's own-image share from 99.93% to 7.88%; companion evidence overturns its cat-versus-truck margin. Across uniform contexts, mean direct self-weight falls from 99.85% to 3.41%; the equilibrium smoother retains 8.15% self-weight. Centered squared dispersion falls to 0.444% of its input value. Of 1,536 query–context placements, 829 initially correct predictions become wrong and 27 initially wrong predictions become correct. Every loss has positive own-image evidence outweighed by companions. Figure 8 in Appendix A.4.3 shows the kernel matrices and population margins; the interventions test the explanation beyond this identity.

The identity also gives a positive criterion for retaining a decision. For an initially correct query of class y, define $b_{ik} = \max_{j \neq i}[Z_{0,jk} - Z_{0,jy}]_+$, where $[u]_+ = \max(u, 0)$. Its class remains correct whenever, for every competitor k,

$$(S_*)_{ii}(Z_{0,iy} - Z_{0,ik}) > (1 - (S_*)_{ii}) b_{ik}. \tag{10}$$

The right side bounds adverse companion evidence. The guarantee requires task margins and context information beyond stochasticity or contraction.

Remove the context term, recover the prediction. For uniform weights $U = \mathbf{1}\mathbf{1}^\top/N$, the exact smoother is $S_U = \beta I + \alpha U$. Thus $x^*_i = \beta a_i + \alpha \bar{a}$ and centered squared dispersion retains $\beta^2 = 0.25\%$ at the focal setting. This suggests the label-free correction

$$\widetilde{X} = \frac{X^* - \alpha \mathbf{1}\bar{a}^\top}{\beta}, \qquad \bar{a} = \frac{1}{N}\sum_j a_j. \tag{11}$$

With the original head, corrected full-dimensional accuracy is 87.96%, a difference of -0.13 pp from the input ($[-0.78, 0.46]$); nearest-context accuracy recovers to 87.30%. Together, the frozen-weight and context-removal controls identify accessible class evidence obscured by adaptive mixing. The correction uses the original context mean and amplifies errors by $1/\beta$; its present value is diagnostic, with no general reconstruction guarantee.

Complementary readout controls use the preceding 16D feature bank. Across three complete 10,000-image test partitions, input accuracy is 75.33%, adaptive accuracy 29.41%, and accuracy after refitting an affine head on refined training examples 29.42%. Context correction reaches 73.93% with the original head (Appendix A.4.6). A shared per-image head cannot generally subtract an offset that varies with the companion set; the correction explicitly supplies that information.

Why initially selective weights can become diffuse. Evaluate the dot-product logits at $X_U = \beta A + \alpha \mathbf{1}\bar{a}^\top$. Up to a rowwise constant,

$$\ell_{ij}(X_U) = s\beta^2\langle a_i, a_j\rangle + s\alpha\beta\langle \bar{a}, a_j\rangle. \tag{12}$$

The pairwise coefficient falls from 30 to 0.075, while the second term favors the same keys for every query. Builder-only interventions support the role of weakened pairwise distinctions. In twelve 16D confirmation contexts, centering the builder inputs barely changes accuracy (34.38% to 34.90%). Raising sharpness to $30/\beta^2 = 12{,}000$ restores the 76.56% input accuracy, with direct self-weight 99.74%; frozen weights recover equally well (Appendix A.4.8). This is recovery through nearly identity mixing. Equation (12) describes the near-uniform regime, not a universal basin or transition criterion.

All focal trajectories meet residual tolerance 10^{-13} and 100-step movement tolerance 10^{-11}. The full study retains one unresolved non-focal solve among 16,384; its capped prediction remains in the reported accuracy. Even without an adaptive contraction certificate, residual $r = X - \alpha PX - \beta A$ bounds the value-reconstruction error by $\|X - SA\|_{\infty,2} \le \|r\|_{\infty,2}/\beta$. At the focal setting this is at most 2×10^{-12} (Appendices A.2.7, B.1.4).

5 Finite computation can change outer stability

Suppose another loop uses the refined state to update its input. If $T_K(A)$ comprises K inner updates from $X_0 = g(A)$, then

$$H_0 = Dg(A), \qquad H_{t+1} = J_t H_t + \beta I, \qquad J_t = D_X F_A(X_t), \quad DT_K(A) = H_K. \tag{13}$$

This derivative follows the whole trajectory, whereas (5) differentiates its equilibrium (Grazzi et al., 2020; Scieur et al., 2022). Input initialization gives $H_0 = I$; an input-independent saved state gives $H_0 = 0$.

Fix a reference anchor A_0 and use the same inner initialization rule and budget for both terms in

$$G_K(A) = (1-\zeta)A + \zeta\left[A_0 + C_\delta\left(\gamma[T_K(A) - T_K(A_0)]\right)\right], \tag{14}$$

where $0 < \zeta \le 1$ is damping, $\gamma \ge 0$ is feedback strength, and C_δ projects each row onto the Euclidean ball of radius δ. The subtraction fixes A_0 for every budget: changing K changes its local response without moving the reference point. Clipping bounds displacement and is inactive near A_0, where

$$DG_K(A_0) = (1-\zeta)I + \zeta\gamma H_K. \tag{15}$$

Iterating (13) on a common contraction domain gives

$$\|H_K\| \le q^K\,\|H_0\| + \frac{\beta}{m}(1-q^K), \qquad q = 1-m. \tag{16}$$

This is a convex combination of initializer sensitivity and the equilibrium-response bound. Since $0 < m \le \beta$, an initializer with sensitivity at most one gives $\|H_K\| \le \beta/m$ for every K. Rowwise projection is nonexpansive in the same norm. Consequently

$$\boxed{\gamma < m/\beta} \tag{17}$$

guarantees outer contraction at every budget on a common invariant domain (Appendix A.2.6). The bound applies to a fixed number of updates; it does not differentiate the stopping decision of an adaptive solver. Outside this certificate, the actual finite derivative determines local stability.

The controlled K2 example in Figure 4 has a stable inner equilibrium, yet $K = 16$, $\gamma = 0.1$, and $\zeta = 0.5$ give outer spectral radius 1.374. The first sampled perturbation shrinks by a factor of 0.512; repeated outer updates instead grow its displacement from 10^{-4} to 0.01559. One shrinking direction therefore misses an expanding mode. At $\gamma = 0.5$, the same short solve has spectral radius 4.869, while $K = 128$ gives 0.775 and decay. Central differences validate the full finite derivative to relative error below 1.2×10^{-9}.

Transient amplification occurs when J_t expands before the trajectory settles. A scalar positive-mixer construction has finite response 58.6 at $K = 26$ but equilibrium response 0.15. Its outer map

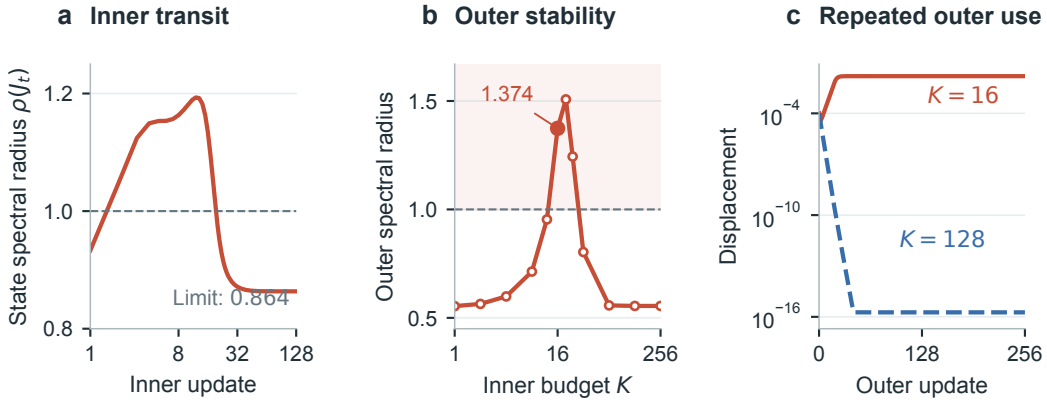


Figure 4: **An intermediate inner budget destabilizes the outer loop.** Duplicate-pair/K2 confirmation example, $\alpha = 0.85, s = 3, N = 16, d = 8$. (a) The inner trajectory passes through states with local spectral radius above one before reaching an attracting equilibrium, $\rho(J) = 0.864$. (b) With outer feedback $\gamma = 0.1$ and damping $\zeta = 0.5$, the full outer derivative has spectral radius 1.374 at $K = 16$ but falls below one at larger budgets. The horizontal boundary is one. (c) At the same $\gamma = 0.1$, the saved outer path grows for $K = 16$ and decays for $K = 128$. The budget sweep and spectral explanation are exploratory; the two outer paths were prespecified.

expands only at $K = 23$–26 in the checked range (Appendix A.5). These examples lie outside the global contraction certificate; within it, (17) controls every budget.

6 What a converged average guarantees

Table 2 separates four questions, their mathematical objects, and the evidence each requires.

Table 2: **Match the guarantee to the quantity being used.** The bounds use the maximum-row norm and a common invariant domain; $\|Dg\| \leq 1$ is required for the all-budget result.

Question	Relevant object	Supported conclusion
Does the solve converge?	$m = \beta - \alpha R L_P$	$m > 0$ guarantees a unique limit; the class threshold is sharp.
Does the limit amplify input changes?	$H = \beta(I - J)^{-1}$	Gain can approach β/m even as the positive frozen map approaches identity.
Are predictions preserved?	Own and companion margins	Equation (10) certifies class preservation; context removal tests recoverability.
Is downstream feedback stable?	The finite derivative H_K	Intermediate budgets can expand; $\gamma < m/\beta$ suffices for every budget.

Scope. Sharpness holds over the stated builder class; kernel-specific thresholds remain open, and global certificates can be conservative. The RED/NLM probe establishes possibility. The classification study isolates cross-image mixing with one pretrained encoder, dataset, and fixed heads. The prevalence and magnitude of this interference in trained architectures remain open. Learned projections, multi-head structure, normalization, and patch-level mixing require their own analysis. The complementary 16D controls have incomplete historical extraction provenance; the documented full-dimensional study has a separate scope.

Positive weights constrain output values; rebuilding introduces another feedback path. Table 2 gives a practical sequence: compute m on a common domain where builder bounds permit; compare adaptive and frozen responses and class margins; examine H_K at the iteration budget actually consumed. Convergence, input response, task fidelity, and downstream stability each require their own evidence.

REPRODUCIBILITY STATEMENT

The separate supplementary PDF contains the worked example, complete proofs, supporting experiments, and controls in Appendix A; preprocessing, experimental grids, numerical tolerances, and statistical procedures in Appendix B.1; and provenance and verification scope in Appendix B.2. The full reproducibility package will be made available with the final submission upon acceptance, or earlier through an anonymous link shared with reviewers and area chairs during rebuttal. It will contain prepared inputs and fitted readouts; complete recorded numerical grids, query outcomes, and context memberships; experiment and analysis source; independent numerical checks; the Lean development with its statement-level coverage map; vector figure generators and image provenance; environment specifications and execution instructions; and machine-readable configurations, source identities, hashes, and manifests. External datasets and encoder weights will be identified by their original sources and exact recorded identities.

AI USE STATEMENT

We used ChatGPT 6 Pro, Codex Astra-6, Fable 5.1 and 5.1 Max, and Gemini 3.1 Pro. The authors established the research concept, core mathematical formulation, initial experimental designs, figure concepts, and initial prose. AI feedback helped refine theorem statements, assumptions, notation, experimental qualifications, and interpretation, and identify gaps motivating additional experiments. AI tools substantially assisted proof development and writing, particularly in Appendix A.2, supplying essential steps, strengthening arguments, and developing alternative proof routes from author-provided claims and outlines.

AI assistance also covered implementation, debugging, execution, and testing of experiment, analysis, and plotting code; literature discovery and bibliographic verification; drafting, reorganizing, and substantively rewriting manuscript and supplementary text, including the abstract and captions; simulated peer review; and data preparation, reproducibility documentation, anonymization, and packaging. Image experiments use established datasets and their existing labels. Synthetic test cases use specified mathematical constructions, and figures are rendered by code from those constructions or recorded results. No source images or labels were synthesized by a generative model.

The authors reviewed the final text, mathematics, code, data, figures, and artifacts, and checked references against original sources. Mathematical checks combined author inspection, additional AI review, numerical and exact-arithmetic tests, and Lean checking. Formal coverage is limited to the scope documented in Appendix B.2; the detailed statement-level map will accompany the reproducibility release. The authors checked the correspondence between manuscript mathematics, implementations, and formalizations, and take full responsibility for the final claims and supporting materials.

ETHICS STATEMENT

This work analyzes mathematical feedback systems and uses existing image datasets. It involves no new human-subject data collection. The findings concern controlled refinement procedures and do not establish a safety guarantee for deployed models.

REFERENCES

Claudio Altafini. Multistability of self-attention dynamics in transformers. *IEEE Transactions on Automatic Control*, pp. 1–16, 2026. doi: 10.1109/TAC.2026.3687461. URL https://doi.org/10.1109/TAC.2026.3687461. Early access.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. Deep equilibrium models. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/01386bd6d8e091c2ab4c7c7de644d37b-Abstract.html.

Shaojie Bai, Vladlen Koltun, and J. Zico Kolter. Stabilizing equilibrium models by jacobian regularization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 554–565, 2021. URL https://proceedings.mlr.press/v139/bai21b.html.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Caron_Emerging_Properties_in_Self-Supervised_Vision_Transformers_ICCV_2021_paper.html.

Miguel Á. Carreira-Perpiñán. Fast nonparametric clustering with gaussian blurring mean-shift. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 153–160, 2006. URL https://faculty.ucmerced.edu/mcarreira-perpinan/papers/icml06.pdf.

Valérie Castin, Pierre Ablin, and Gabriel Peyré. How smooth is attention? In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 5817–5840, 2024. URL https://proceedings.mlr.press/v235/castin24a.html.

Benjamin P. Chamberlain, James Rowbottom, Maria Gorinova, Stefan Webb, Emanuele Rossi, and Michael M. Bronstein. GRAND: Graph neural diffusion. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1407–1418, 2021. URL https://proceedings.mlr.press/v139/chamberlain21a.html.

Ming Chen, Zhewei Wei, Zengfeng Huang, Bolin Ding, and Yaliang Li. Simple and deep graph convolutional networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1725–1735, 2020a. URL https://proceedings.mlr.press/v119/chen20v.html.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607, 2020b. URL https://proceedings.mlr.press/v119/chen20j.html.

Wentao Cui and Yuhong Guo. Parameterless transductive feature re-representation for few-shot learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2212–2221, 2021. URL https://proceedings.mlr.press/v139/cui21a.html.

George Dasoulas, Kevin Scaman, and Aladin Virmaux. Lipschitz normalization for self-attention layers with application to graph neural networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2456–2466, 2021. URL https://proceedings.mlr.press/v139/dasoulas21a.html.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: pure attention loses rank doubly exponentially with depth. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2793–2803, 2021. URL https://proceedings.mlr.press/v139/dong21a.html.

Guoji Fu, Peilin Zhao, and Yatao Bian. p-Laplacian based graph neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 6878–6917, 2022. URL https://proceedings.mlr.press/v162/fu22e.html.

Johannes Gasteiger, Aleksandar Bojchevski, and Stephan Günnemann. Predict then propagate: Graph neural networks meet personalized PageRank. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=H1gL-2A9Ym.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. The emergence of clusters in self-attention dynamics. In *Advances in Neural Information Processing Systems*, volume 36, pp. 57026–57037, 2023. doi: 10.52202/075280-2493. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/b2b3e1d9840eba17ad9bbf073e009afe-Abstract-Conference.html.

David F. Gleich, Lek-Heng Lim, and Yongyang Yu. Multilinear PageRank. *SIAM Journal on Matrix Analysis and Applications*, 36(4):1507–1541, 2015. doi: 10.1137/140985160. URL https://www.stat.uchicago.edu/~lekheng/work/mpr.pdf.

Riccardo Grazzi, Luca Franceschi, Massimiliano Pontil, and Saverio Salzo. On the iteration complexity of hypergradient computation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3748–3758, 2020. URL https://proceedings.mlr.press/v119/grazzi20a.html.

Fangda Gu, Heng Chang, Wenwu Zhu, Somayeh Sojoudi, and Laurent El Ghaoui. Implicit graph neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 11984–11995, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/8b5c8441a8ff8e151b191c53c1842a38-Abstract.html.

Hyunjik Kim, George Papamakarios, and Andriy Mnih. The Lipschitz constant of self-attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5562–5571, 2021. URL https://proceedings.mlr.press/v139/kim21i.html.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Tam Nguyen, Tan M. Nguyen, and Richard G. Baraniuk. Mitigating over-smoothing in transformers via regularized nonlocal functionals. In *Advances in Neural Information Processing Systems*, volume 36, pp. 80233–80256, 2023. doi: 10.52202/075280-3516. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/fde1a69a5b6e554b2f1f727197d2651d-Abstract.html.

Kenta Oono and Taiji Suzuki. Graph neural networks exponentially lose expressive power for node classification. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1ldO2EFPr.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, David Kreil, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=tL89RnzIiCd.

Edward T. Reehorst and Philip Schniter. Regularization by denoising: Clarifications and new interpretations. *IEEE Transactions on Computational Imaging*, 5(1):52–67, 2019. doi: 10.1109/TCI.2018.2880326. URL https://phil-schniter.web.app/pdf/tci19_red.pdf.

Yaniv Romano, Michael Elad, and Peyman Milanfar. The little engine that could: Regularization by denoising (RED). *SIAM Journal on Imaging Sciences*, 10(4):1804–1844, 2017. doi: 10.1137/16M1102884. URL https://elad.cs.technion.ac.il/wp-content/uploads/2018/02/LittleEngine-Final.pdf.

Damien Scieur, Quentin Bertrand, Gauthier Gidel, and Fabian Pedregosa. The curse of unrolling: Rate of differentiating through optimization. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17133–17145, 2022. doi: 10.52202/068431-1246. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6d53193a098b982229340a7c3eb0ecbf-Abstract-Conference.html.

Matthew Thorpe, Hedi Xia, Tan Nguyen, Thomas Strohmer, Andrea L. Bertozzi, Stanley J. Osher, and Bao Wang. GRAND++: Graph neural diffusion with a source term. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=EMxu-dzvJk.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJXMpikCZ.

James Vuckovic, Aristide Baratin, and Remi Tachet des Combes. On the regularity of attention. *arXiv preprint arXiv:2102.05628*, 2021. URL https://arxiv.org/abs/2102.05628.

Ezra Winston and J. Zico Kolter. Monotone operator equilibrium networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 10718–10728, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/798d1c2813cbdf8bcdb388db0e32d496-Abstract.html.

Xinyi Wu, Amir Ajorlou, Zihui Wu, and Ali Jadbabaie. Demystifying oversmoothing in attention-based graph neural networks. In *Advances in Neural Information Processing Systems*, volume 36, pp. 35084–35106, 2023. doi: 10.52202/075280-1524. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6e4cdfdd909ea4e34bfc85a12774cba0-Abstract-Conference.html.

Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems*, volume 16, pp. 321–328, 2004. URL https://proceedings.neurips.cc/paper_files/paper/2003/hash/87682805257e619d49b8e0dfdc14affa-Abstract.html.

Jiong Zhu, Yujun Yan, Lingxiao Zhao, Mark Heimann, Leman Akoglu, and Danai Koutra. Beyond homophily in graph neural networks: Current limitations and effective designs. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7793–7804, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/58ae23d878a47004366189884c2f8440-Abstract.html.