CONDITIONED DIRECT FEEDBACK ALIGNMENT VIA ACTIVITY AND ERROR GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

Direct feedback alignment (DFA) trains hidden layers with fixed random projections of the output error. Even when this feedback provides useful credit, unequal scales across activity or error directions can distort the local update. We study conditioned DFA (nDFA), a family that adapts established inverse-moment preconditioning to either side of this update. An aligned linear analysis describes how activity conditioning changes spectral learning rates and early-stopping risk. Synthetic experiments show gains from both stabilization and changes in update direction. Confirmatory CIFAR-10 experiments show that activity conditioning, including its FOOF formulation, improves tuned DFA at matched measured training time. Controls support a role for centered correlations beyond the tested mean and diagonal alternatives, while early-only conditioning retains most of the benefit with less training work. Conditioning also benefits backpropagation, and timing interventions do not establish a mechanism specific to random feedback. Error conditioning improves short-budget models and changes update directions, but its small additional improvement under stable training does not survive correction for multiple comparisons. Predicted benefits on image-background benchmarks are not confirmed. These findings establish practical benefits and important limits of conditioning learning with fixed random feedback.

1 INTRODUCTION

Credit assignment asks how a synapse should change when an outcome depends on many neurons. Backpropagation (BP) assigns this credit by propagating output errors through the transpose of each forward weight matrix. Feedback alignment (FA) and direct feedback alignment (DFA) relax this requirement by replacing transported errors with fixed random feedback signals (Lillicrap et al., 2016; Nøkland, 2016). These rules connect to biological credit-assignment models, including segregated dendritic compartments (Lillicrap et al., 2020; Guerguiev et al., 2017; Sacramento et al., 2018). Their learning can nevertheless be brittle: their success depends on whether the random feedback produces a usable hidden-layer update (Bartunov et al., 2018; Launay et al., 2020; Garg and Vempala, 2022).

Many analyses of DFA focus on the feedback pathway: its rank, scale, alignment, and compatibility with the network architecture. Here we study a complementary failure mode that can occur even when feedback provides a useful descent component: the local DFA update is an outer product with the presynaptic activity, so high-variance activity directions that carry no task information can dominate the update with nuisance structure rather than useful credit.

We ask whether reweighting directions in the local activity and error signals improves learning while leaving feedback fixed. We call this reweighting conditioning. Anisotropy means unequal second-moment energy across directions. Activity anisotropy weights combinations of presynaptic inputs; error anisotropy weights combinations of the teaching signals delivered to hidden units. Conditioning can act on either side of the local outer product.

Our contribution is a controlled study of established conditioning operators under imperfect fixed feedback. The activity operator is the FOOF update applied to DFA errors (Benzing, 2022); its matrix inverse is not new. We separate activity and error factors, test which activity statistics matter, and compare conditioning schedules against tuned baselines at matched measured work. A joint confirmation supports activity conditioning and an early-only efficiency tradeoff, but limits

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.

the additional error-side benefit and the transfer of synthetic nuisance predictions. Distinguishing stabilization, faster optimization and improved generalization is central to these conclusions.

2 CONDITIONED DFA

We now turn the geometric failure mode into a learning rule. For layer ℓ, let W_ℓ be its weight matrix, $h_{\ell-1}$ its input activity, a_ℓ its preactivation, and ϕ its nonlinearity. The output error e is the derivative of the per-example loss with respect to the output preactivation. A fixed matrix B_ℓ maps this error to the hidden-layer dimension. For an elementwise nonlinearity, DFA uses

$$\delta_\ell^{\mathrm{DFA}} = (B_\ell e) \odot \phi'(a_\ell), \qquad G_\ell^{\mathrm{DFA}} = \delta_\ell^{\mathrm{DFA}} h_{\ell-1}^\top. \tag{1}$$

Here $\odot$ denotes elementwise multiplication and $\eta > 0$ the learning rate. The raw step is $-\eta G_\ell^{\mathrm{DFA}}$; minibatches average these outer products. Batch-normalized layers also apply the local normalization derivative.

Conditioning the local outer product. The minibatch update combines local errors with presynaptic activities. To condition either factor, we estimate the uncentered moments

$$C_{A,\ell} = \mathbb{E}[h_{\ell-1} h_{\ell-1}^\top], \qquad C_{E,\ell} = \mathbb{E}[\delta_\ell^{\mathrm{DFA}} \delta_\ell^{\mathrm{DFA}\top}], \tag{2}$$

where the expectation averages examples, including means as well as fluctuations. Separate damping $\lambda_A, \lambda_E > 0$ stabilizes each inverse.[1] With I denoting the identity matrix, define

$$P_{A,\ell} = (C_{A,\ell} + \lambda_A I)^{-1}, \qquad P_{E,\ell} = (C_{E,\ell} + \lambda_E I)^{-1}.$$

The two-factor conditioned-DFA family is

$$\begin{aligned} \Delta W_\ell^{A\text{-nDFA}} &= -\eta\, G_\ell^{\mathrm{DFA}} P_{A,\ell}, \\ \Delta W_\ell^{E\text{-nDFA}} &= -\eta\, P_{E,\ell} G_\ell^{\mathrm{DFA}}, \\ \Delta W_\ell^{K\text{-nDFA}} &= -\eta\, P_{E,\ell} G_\ell^{\mathrm{DFA}} P_{A,\ell}. \end{aligned} \tag{3}$$

For the norm-matched comparisons, write the selected matrix above as U_ℓ before multiplication by $-\eta$ and use

$$\widetilde{G}_\ell = \begin{cases} \|G_\ell^{\mathrm{DFA}}\|_F\, U_\ell / \|U_\ell\|_F, & \|U_\ell\|_F > 0, \\ 0, & \|U_\ell\|_F = 0. \end{cases} \tag{4}$$

This operation applies only to hidden weight matrices. Bias, classifier and normalization parameters retain the underlying learner's updates. We use "nDFA" without a prefix for the activity-only rule, and "K-nDFA" for the two-sided Kronecker form; A, E, and K abbreviate these three rules. The "n" denotes inverse-second-moment normalization, not an exact natural-gradient method. All three updates are block-local after output-error broadcast: they require minibatch-wide layer statistics and damped solves, but not backpropagated hidden errors. They are not synaptically local. A direct dense solve requires $O(d^2)$ state and $O(d^3)$ factorization. When a minibatch has fewer examples than units, exact sample-space identities replace this solve by a smaller Gram-system solve without changing the update (Appendix C.3); K-nDFA still applies both factors. The activity rule shares FOOF's core operator (Benzing, 2022); here the error source is DFA and moment estimation and norm matching follow the stated protocol. The two-sided form resembles KFAC, but a DFA pseudo-error moment is not a Fisher factor.

Candidate sources of benefit. Activity conditioning is relevant when high-variance presynaptic directions are nuisance-dominated. Error conditioning targets a different pathology: unequal second-moment energy across local credit coordinates, which can arise from class imbalance, unequal residual scales, saturation/gating, or anisotropic random-feedback routing even when the input is not noisy. Combining the factors may help when each improves the task-relevant part of the update. Estimating C_E is especially fragile because gated DFA errors can be low-rank; under-damped inversion can amplify poorly estimated directions. The rule uses layer statistics without nuisance labels or a known task subspace. Equal algebraic roles do not imply equal damping or robustness.

[1] Terminology: except when explicitly discussing the power-$1/2$ ZCA (zero-phase component analysis) whitening control, "inverse-second-moment preconditioning" refers to the power-1 updates below. "Activity/input side" refers to right multiplication in weight-gradient coordinates; "error/output side" refers to left multiplication.

Error scaling and update magnitude. Error moments use per-example signals. A local error computed from a mean minibatch loss is multiplied by the batch size before forming its moment. The factor-separated comparisons use these per-example moments, independently validation-selected damping, and the layerwise norm matching in Eq. 4.

2.1 Linearized analysis: inverse-second-moment conditioning

To isolate the input-covariance effect, take Gaussian inputs $x \sim \mathcal{N}(0, \Sigma)$ with $\Sigma \succ 0$, target $W^\star x$, hidden and output weights $W^{(1)}, W^{(2)}$, and fixed feedback B. The residual operator is $\varepsilon = W^{(2)}W^{(1)} - W^\star$; here the presynaptic second moment $C_{\ell-1}$ of Eq. 2 specializes to the input covariance Σ. The expected first-layer updates are $g^{\mathrm{BP}} = -(W^{(2)})^\top \varepsilon \Sigma$, $g^{\mathrm{DFA}} = -B\varepsilon\Sigma$, and $g^{\mathrm{nDFA}} = -B\varepsilon\Sigma(\Sigma + \lambda_A I)^{-1}$. Here g denotes a descending update per unit learning rate, whereas G in Eq. 1 is gradient-like: the corresponding direction is $g = -G$. The linear-Gaussian loss Hessian has Kronecker form $(W^{(2)\top}W^{(2)}) \otimes \Sigma$. Right-multiplication by $(\Sigma + \lambda_A I)^{-1}$ is the damped input-side inverse-second-moment factor used by Kronecker-factored approximate curvature (KFAC) and related Newton or natural-gradient approximations (Amari, 1998; Martens and Grosse, 2015). In the nonlinear experiments below, we use this curvature-motivated preconditioner on the DFA update. The analogous factorization motivates separate left and right inverses, but is not an exact Fisher identity: DFA errors depend on random feedback, gating and residuals (Appendix B).

Spectral prediction. In the post-alignment linear-Gaussian regime with population second moments and damping $\lambda_A > 0$, let λ_i be the eigenvalues of Σ. nDFA replaces the per-eigendirection input factor λ_i by $\lambda_i/(\lambda_i + \lambda_A)$. The associated input-side condition number is

$$\kappa_{\mathrm{nDFA}} = \kappa(\Sigma)\,\frac{\lambda_{\min} + \lambda_A}{\lambda_{\max} + \lambda_A},$$

where $\kappa(\Sigma) = \lambda_{\max}/\lambda_{\min}$ and the extrema refer to the input eigenvalues (Fig. 1B–C). Realized learning also depends on the residual, initialization, step size, sampling noise and task orientation. Our empirical hypothesis is narrower: nDFA should help most when otherwise active high-variance directions are task-irrelevant nuisance, should be redundant under isotropy, and should offer little advantage over a well-tuned exact-gradient method when high-variance directions carry the task. The analogous left-side prediction is conditional: error nDFA may help when high-variance local-error directions obscure useful credit, but can hurt when those directions carry the task or when C_E is poorly estimated or under-damped. Figure 1A illustrates the geometry; panel D shows why the realized gain depends on task orientation even at the same condition number.

Appendix B gives the exact input-factor identity and alignment assumptions. Its separate two-block regression calculation illustrates how re-timing task and nuisance fitting changes early-stopping risk, a mechanism studied previously for preconditioned and differently scaled gradient dynamics (Amari et al., 2021; Heckel and Yilmaz, 2021). It does not explain the acquisition of DFA alignment or establish a new general principle of early stopping.

A complementary construction holds the raw update fixed while varying activity and error orientations (Appendix K). It isolates how either factor can improve or worsen approximate credit at the same moment spectrum (Appendix Fig. 11A–C), extending the geometric question beyond the aligned input-side model.

3 Experimental setup

The experiments distinguish directional effects from stabilization, then test whether those effects remain useful with stronger baselines. An eight-class synthetic suite varies task geometry, label noise and sample size. Historical MNIST/Fashion-MNIST (Xiao et al., 2017) studies separate the activity and error factors at short training horizons. Their protocols and test-selection limitations remain explicit (Appendices D–E).

The primary practical study uses CIFAR-10 (Krizhevsky, 2009) ReLU MLPs with batch normalization (BN) in the core arms, per-method validation selection and equal measured learning time on the same accelerator. It compares tuned DFA and BP, forward-decorrelated DFA (FD-DFA), activity conditioning, early-only activity conditioning, and FOOF with either credit rule. FOOF uses averaged

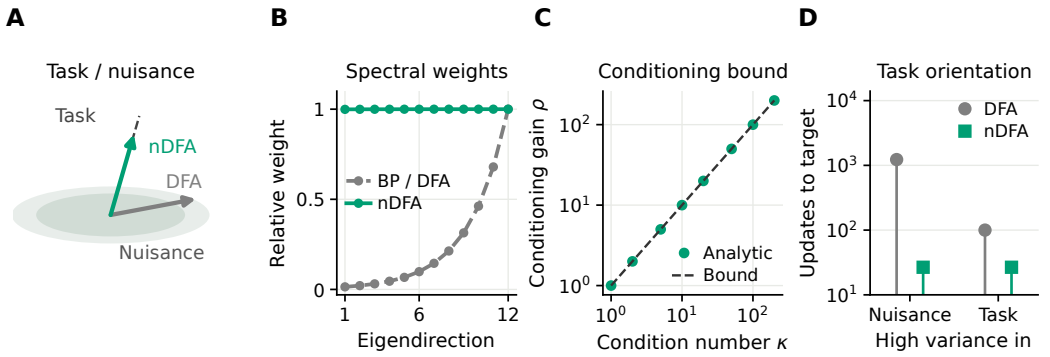

Figure 1: **Conditioning reweights a local update. A,** A mixed residual (dashed direction) has task and nuisance components. An anisotropic activity moment weights its nuisance component strongly; inverse-moment conditioning reduces that imbalance. **B,** In the aligned linear model, raw BP/DFA weight eigendirections by λ_i, while activity conditioning gives $\lambda_i/(\lambda_i+\lambda_A)$ (relative weights shown). **C,** The analytic gain in input conditioning, $\rho = \kappa(\Sigma)/\kappa_{\mathrm{nDFA}}$, approaches its bound $\kappa(\Sigma)$ under weak damping. Panels B–C use $\lambda_A = 10^{-3}$ and $\lambda_{\min} = 1$. **D,** Population linear simulations at condition number 50, with separately selected learning rates: updates to a target loss depend on task orientation (protocol in Appendix B).

activity moments and periodic inverse updates without norm matching; A uses current minibatch moments and Eq. 4. Separate interventions examine centered covariance, conditioning windows at constant learning rate, and temporary error conditioning. Image-background tasks test the proposed transfer from synthetic nuisance geometry to real images.

After development, a joint protocol fixes settings, endpoints and ten paired training seeds. Work-budget studies select the checkpoint with lowest validation cross-entropy (CE); fixed-epoch interventions use the final checkpoint. All confirmation training completes before joint test evaluation. The standard CIFAR/MNIST/Fashion test sets had been used in earlier project studies; new seeds do not make these datasets previously unseen. The primary family contains 21 paired tests with global Holm correction. Intervals are individual Student t intervals over training seeds, conditional on the data split and selection procedure. Appendix A reports every contrast, protocol deviation and secondary analysis. Historical experiments retain their original replication units and analysis families.

4 Results

4.1 Synthetic controls separate stabilization from geometry

The synthetic suite tests the predicted dependence on whether high-variance directions carry task signal or nuisance (Fig. 2A–C; per-method results in Table 20).

The norm-matched controls provide the clearest evidence for directional benefit beyond update scale (Fig. 4A). The broader shared-rate sweep also strongly favors nDFA, but raw DFA has extreme finite training losses in nearly all stressed runs. Its large accuracy gaps therefore combine stabilization with any change in useful credit (Table 21). All five global seeds favor nDFA after averaging the fixed grid. These comparisons do not establish an advantage over individually tuned, near-converged learners.

The focused factor ablation shows an activity-side benefit in the nuisance-dominant and low-sample/noisy cells (Fig. 2D).

In the short-horizon task-aligned sweep, conditioning improves DFA but tuned BP remains stronger. The focused 100-epoch clean, norm-matched control shows no activity benefit (Table 25). Thus the broad short-horizon gain should not be read as a sustained generalization advantage when high-variance directions carry task signal.

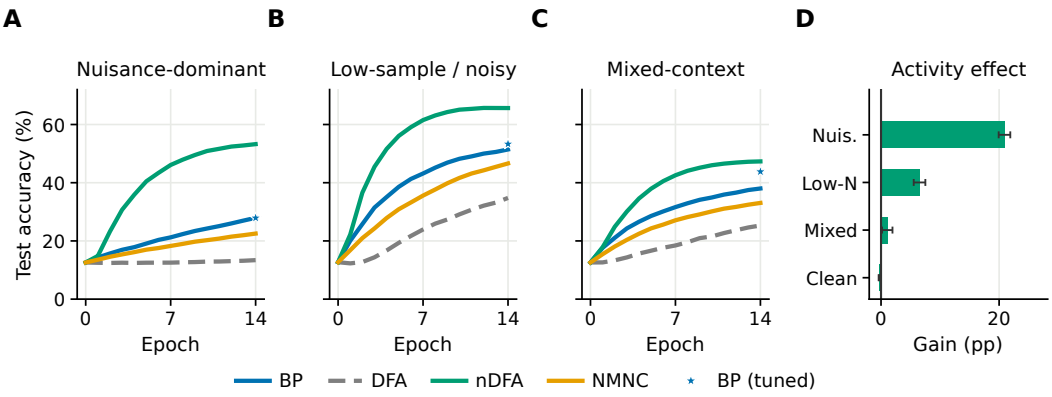

Figure 2: **Conditioned DFA and the activity-side regime. A–C,** Fixed-full-rank nDFA separates from raw DFA in the three stressed regimes; stars mark BP endpoints from a separate per-regime learning-rate search selected on test accuracy. Raw DFA's shared-rate high-loss instability contributes to the displayed gaps. The 14-epoch curves average the fixed designed grid; inferential bands are omitted. NMNC: neural-manifold noise correlation (Kang et al., 2026), a perturbation-based local rule. **D,** Activity-only nDFA minus DFA after 100 epochs in the focused nuisance, low-sample/noisy, mixed and clean task-aligned cells: $+20.9$, $+6.5$, $+1.1$ and -0.3 percentage points (pp), respectively. Bars show mean $\pm$ SEM over five data seeds, each averaging three feedback seeds.

In noisy-label vision MLPs, conditioning lifts raw DFA close to BP on Fashion-MNIST and CIFAR-10 (Table 29). These sweeps use fixed full-rank feedback. Their shared-rate, short-horizon comparisons do not establish parity with individually tuned BP.

On a ColoredMNIST-style task (Arjovsky et al., 2019), nDFA has lower seed variability than DFA and approaches BP under a reversed color–label correlation. The paired accuracy gain does not survive Holm correction at eight runs, supporting a descriptive stability result rather than superiority (Appendix D.6).

Across cells where raw DFA learns, conditioning also reduces variation across compound feedback/order seeds (Appendix D.4). This is a descriptive stability observation: feedback and minibatch order vary together. The early alignment trajectories do not isolate an alignment-rescue mechanism, and the controlled timing tests below do not establish DFA specificity.

4.2 Error conditioning: short-budget gains and a stable-training limit

The left factor changes a different part of the local update, so an activity benefit does not establish an error benefit. Historical tanh MNIST and Fashion-MNIST experiments select the two damping values independently and combine them without a joint search. Error conditioning improves raw DFA, and adding it to A improves the short-budget endpoint (Fig. 3). A ReLU/softmax MNIST replication preserves these effects. All conditioned hidden matrices are norm-matched to raw DFA. These fixed-rate studies test early optimization, not superiority to tuned, converged baselines. The original MNIST seeds had already been evaluated on test data before validation development; the separate Fashion and ReLU protocols and later post-hoc extensions are distinguished in Appendix E.

The stable-training CIFAR-10 confirmation gives a more limited result. Adding E during the first quarter of training changes update direction substantially and has small favorable mean accuracy and loss effects, but neither survives the global correction. It also fails the required performance comparisons with diagonal E. Thus it fails the prespecified five-part criterion for an additional error-side benefit. The complete result and scalar-control identity are reported in Appendix A.8; the failure is not evidence that E is exactly ineffective.

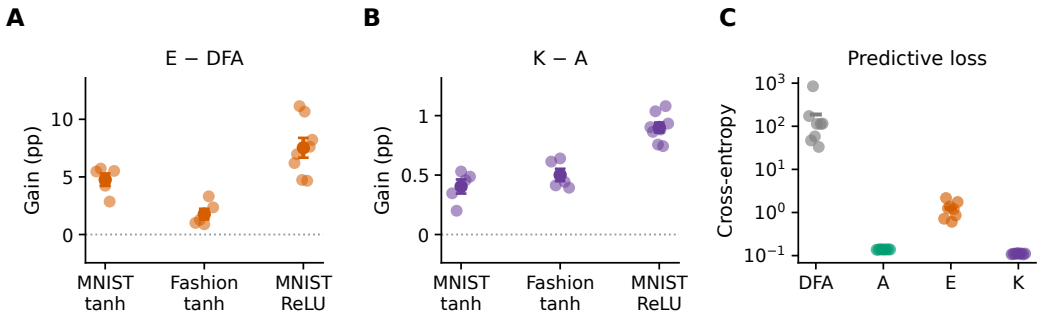

Figure 3: **Historical short-budget error-side gains. A,** Error nDFA improves raw DFA in both tanh confirmations and the ReLU/softmax confirmation. **B,** Adding the error factor after activity conditioning improves all seed-level pairs in all three settings. Dots are model/data-order seeds after averaging three feedback seeds; large points are mean ± SEM. **C,** Final test cross-entropy in the ReLU/softmax confirmation (log scale); dots show the eight seed means and horizontal marks their average. These fixed-rate endpoints do not establish a stable-training advantage; see Appendix A.8.

Rule	Test accuracy (%)	Test CE
DFA (AdamW)	65.66 ± 0.25	0.991 ± 0.003
BP (AdamW)	66.88 ± 0.25	0.963 ± 0.006
FD-DFA	67.01 ± 0.25	0.960 ± 0.006
Activity DFA	67.92 ± 0.27	0.943 ± 0.004
Early activity DFA	68.41 ± 0.34	0.945 ± 0.006
FOOF-DFA	69.21 ± 0.30	0.954 ± 0.007
FOOF-BP	70.03 ± 1.17	0.954 ± 0.008

Table 1: **Joint CIFAR-10 confirmation against tuned baselines.** Mean ± sample SD over ten paired training seeds. Each run receives 200 seconds of synchronized learning work on an H100; the checkpoint is selected by validation CE. Calibration, conditioning and inverse refresh are charged; validation and export are excluded. The 3072–1024–512–10 MLP and data are shared. Core arms use BN; FD-DFA uses forward decorrelation instead. Settings are selected separately. Appendix A gives all arms, estimated FLOPs, update counts and corrected paired tests.

4.3 Controls identify the source of the gain

We next test regularization, scalar rescaling and diagonal adaptation as alternative explanations. On the nuisance-dominant synthetic cell, regularization and early stopping provide much smaller BP gains than activity conditioning (Table 24). Matching raw DFA's hidden gradient norms to BP does not recover the benefit. On noisy-label Fashion-MNIST, however, this scalar correction reaches slightly higher mean accuracy than nDFA (Table 24): scale can suffice when the local direction is already useful. BP norms are used only in this diagnostic; the factor confirmations match raw DFA.

Applying the same BP-norm matching *after* conditioning preserves the activity-side advantage in all three hard cells. The supported effect is therefore directional reweighting beyond a scalar step-size change (Fig. 4A). Across the designed cells, the gain also grows with nuisance energy relative to task energy (Fig. 4B); this is a descriptive association, not an independently validated predictor.

Is BatchNorm already the fix? BatchNorm standardizes per-unit activity (Ioffe and Szegedy, 2015) and recovers much of the gain (Table 28). It matches or exceeds nDFA in the low-sample/noisy and mixed regimes, while the remaining nDFA advantage is concentrated in nuisance-dominant cells. Adding activity conditioning to DFA+BN helps in the nuisance and low-sample regimes but hurts in the mixed and task-aligned regimes (Table 28).

Inverse versus inverse-square-root conditioning. To distinguish inverse-moment conditioning from whitening (Desjardins et al., 2015; Huang et al., 2018), a matched control on the same 128-cell

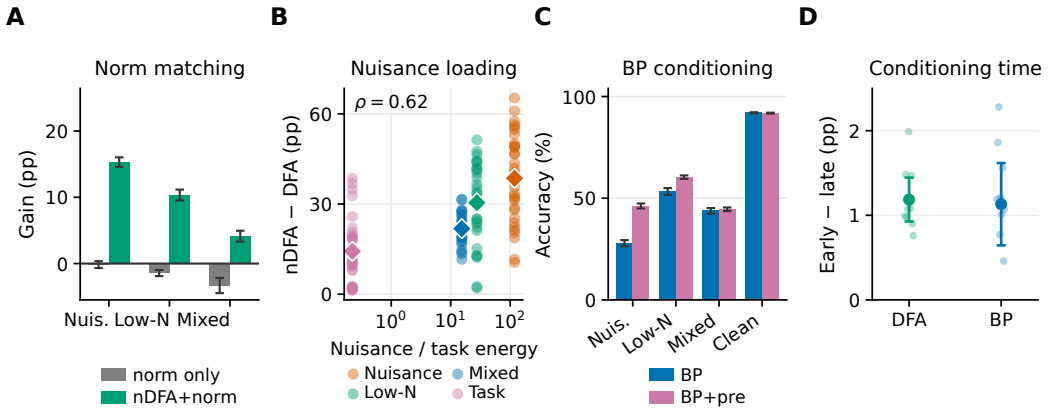

Figure 4: **Directional controls and conditioning time. A,** Activity gains survive instantaneous BP-norm matching of hidden updates (five synthetic training seeds, mean ± SEM). **B,** Gain versus designed nuisance/task energy in the exploratory rank-selected sweep (diamonds: regime means; descriptive). **C,** The same activity operator also helps BP in stressed synthetic conditions (mean ± SEM over five global seeds after test-based learning-rate selection; descriptive). **D,** Early-minus-late conditioning at constant learning rate helps both DFA and BP on CIFAR-10; their interaction is not significant. Dots show ten paired training seeds and bars individual 95% t intervals. Panels A–C are historical controls; D is the joint confirmation.

suite preconditions the DFA update by $(C_{\ell-1} + \lambda I)^{-1/2}$ instead of nDFA's full inverse (Table 26). Inverse-square-root conditioning captures most of the input-side gain. The full inverse adds more in the nuisance-dominant and low-sample/noisy regimes than in the mixed-context or clean task-aligned controls. This is consistent with the spectral motivation, but all comparisons come from the same designed suite.

Centered correlations beyond mean activity. Full uncentered moments combine covariance and the activity mean: $C_A = \mathrm{Cov}(h) + \mu\mu^\top$, where $\mu = \mathbb{E}[h]$. A full-versus-diagonal comparison alone cannot separate these contributions. We therefore compare full and centered conditioning with a control that preserves the mean outer product but removes off-diagonal covariance, $\mathrm{diag}(\mathrm{Cov}(h)) + \mu\mu^\top$. Both richer operators outperform this control in every confirmation seed, with corrected evidence for both contrasts (Fig. 5B). This supports a useful role for centered correlations under the tested recipe, beyond a mean-only explanation; it does not exclude other coordinatewise methods. Hidden-weight Adam also fails to reproduce the synthetic activity gain in the exploratory settings (Kingma and Ba, 2015) (Table 27).

Input reconditioning in exact BP. The same right preconditioner also improves BP in the nuisance-dominant and low-sample/noisy regimes, with little change in the mixed and clean controls (Fig. 4C), after separate tuning over the same learning-rate grid on the reported test data. This is an exploratory control. This supports an activity-geometry mechanism shared with ordinary preconditioning, rather than a DFA-specific theoretical law. nDFA remains higher in the nuisance regime, but the tuning and regularization contributions to that difference are unresolved (Appendix Table 26). A separate validation-selected CIFAR-10 protocol also favors conditioned BP, while changing the ordering of DFA variants (Appendix J).

4.4 Activity gains survive stronger baselines and matched work

Practical value requires gains that remain after tuning the baseline and charging for conditioning. In the ten-seed CIFAR-10 confirmation, activity conditioning, its early-only variant and FOOF-DFA all improve tuned AdamW DFA in accuracy, without increasing mean CE (Table 1; Fig. 5A). Every paired seed favors each method, and all three pass the prespecified practical and corrected statistical criteria. FOOF-DFA has the highest mean accuracy among the tested DFA methods; activity conditioning has lower mean CE. FOOF-BP remains the stronger accuracy reference. These

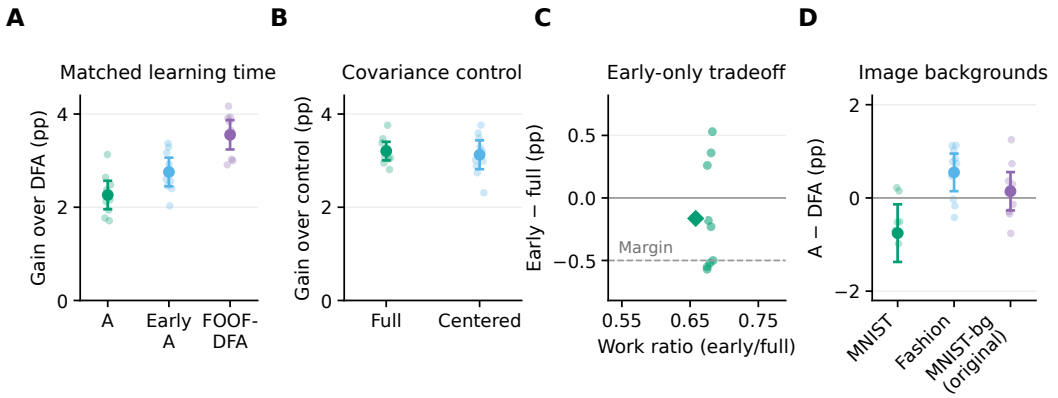

Figure 5: **Confirmed benefits and their limits. A,** Activity, early activity and FOOF-DFA improve tuned DFA at matched measured learning time on CIFAR-10. **B,** Full and centered conditioning outperform the mean-preserving diagonal-covariance control. **C,** In a separate fixed-epoch cohort, early conditioning retains accuracy within the prespecified margin while reducing measured learning work. **D,** Activity-minus-DFA accuracy effects on constructed MNIST/Fashion backgrounds and the original MNIST-background-images task do not establish the predicted gains. Points show ten paired seed contrasts; bars show means and individual 95% t intervals. C shows seed-level work ratios and accuracy differences, with the mean marked separately. See Appendix A for the global correction, loss endpoints, hardware and noninferiority criterion.

tests establish gains over tuned DFA, not superiority to every BP or decorrelation method. The credit-rule-by-FOOF interaction does not establish a preferential benefit for DFA. FOOF also uses substantially more estimated arithmetic work, so matching learning time does not establish a FLOP advantage.

When is conditioning needed? At constant learning rate, conditioning during the first quarter of training outperforms conditioning during the last quarter, but the corresponding interaction does not establish a difference between BP and DFA (Fig. 4D). Shared-checkpoint development interventions also show that switching conditioning on late can improve both credit rules. Thus the evidence supports an early-window advantage, not irreversible nuisance capture or a DFA-specific timing mechanism. In a separate fixed-epoch cosine-schedule experiment, early-only activity conditioning meets the prespecified accuracy noninferiority criterion and uses about one-third less measured learning work than persistent conditioning (Fig. 5C). This tradeoff is conditional on that recipe and differs from the time-matched comparison in Table 1.

Where does the nuisance prediction transfer? The synthetic model motivates a directional prediction, not a guarantee for any noisy input. We test clean images, training-label noise, and images mixed with independent backgrounds on MNIST and Fashion-MNIST, plus the original MNIST-background-images benchmark. The predicted background gains are not confirmed (Fig. 5D). In both constructed background tasks, activity conditioning worsens CE after correction, despite mixed accuracy effects. Input nuisance alone therefore does not identify a beneficial conditioning regime.

Normalization and scope. Earlier width and work studies support activity benefits but show that retuned error damping can nearly suppress E (Appendices F–I). A separate unnormalized protocol produces extreme predictive losses for several conditioned rules. Under a fixed-recipe intervention, BN restores ordinary loss scales for both raw and conditioned DFA; reducing feedback scale alone does not (Appendix J). These controls explain why both accuracy and loss, and the normalization recipe, are necessary to interpret a conditioning benefit.

5 Related Work

Conditioned DFA connects fixed feedback with update preconditioning. FA replaces exact weight symmetry with fixed random feedback, and DFA broadcasts the output error directly to hidden layers (Lillicrap et al., 2016; Nøkland, 2016; Refinetti et al., 2021). Later work scaled DFA to Transformers but exposed harder vision limits (Bartunov et al., 2018; Launay et al., 2020; Wang et al., 2026). We condition its local update; GrAPE instead aligns feedback with forward-mode Jacobian estimates and BP anchors (Caillon et al., 2026), a complementary nonlocal axis.

The activity factor uses FOOF's core operator with DFA errors (Benzing, 2022). The two-sided form resembles KFAC, but its local-error moment need not estimate the Fisher matrix (Martens and Grosse, 2015). Related normalization and decorrelation methods act on the forward representation (Desjardins et al., 2015; Huang et al., 2018; 2019). Task-dependent preconditioning benefits and mode-dependent early stopping predate this study (Amari et al., 2021; Heckel and Yilmaz, 2021). Covariance-power and spectral-gradient analyses show task-dependent effects (Yoshida and Nitanda, 2025; Braun et al., 2026); we apply the factors under fixed DFA feedback. Shampoo also applies structured matrix preconditioning, but builds accumulated gradient-derived factors rather than the current-batch activity and local-error moments used here (Gupta et al., 2018). Our contribution is the controlled factor study under imperfect fixed feedback. Update orthogonalization and activity normalization can likewise change effective FA update rank (Boeshertz et al., 2026); their mechanisms overlap with conditioning even though invertible factors preserve algebraic rank.

Alternative feedback and credit mechanisms, and the population-geometry connection, are compared in Appendix L.

6 Discussion and Limitations

Activity conditioning improves learning with random feedback beyond a scalar step-size change. With stronger baselines and a frozen joint confirmation, its practical benefit survives, centered correlations contribute beyond the tested mean-preserving control, and early-only conditioning offers a useful accuracy–work tradeoff. These are properties of established conditioning operators applied to approximate credit, not a new matrix inverse or a general advantage over backpropagation.

The two factors have unequal empirical support. Error conditioning changes directions and improves historical short-budget models, yet its small additional stable-training gain fails the declared corrected criterion. This outcome is consistent with the importance of the activity factor in FOOF/KFAC comparisons (Benzing, 2022), without implying that DFA's empirical error moment is a Fisher factor or that error conditioning is universally ineffective. Likewise, the failed image-background predictions show that the synthetic geometry cannot be inferred from a dataset's noise label. Our aligned theory does not explain alignment acquisition, and the timing interventions do not establish a DFA-specific mechanism.

The main practical evidence concerns MLPs and fixed data splits. Training-seed uncertainty does not include dataset or hyperparameter-selection uncertainty; standard test sets had historical exposure, and one benchmark cohort changed accelerator before confirmation. Equal measured learning time excludes validation and export and does not imply equal FLOPs. Exact sample-space solves reduce the cost of wide layers, but moments and solves remain layer-wide operations, not synaptically local plasticity. Within these limits, activity geometry provides a reproducible way to improve the use of fixed random feedback.

Reproducibility Statement

The appendix is self-contained and gives the protocols, selected settings, all primary confirmation contrasts and their seed-level values, assumptions, and deviations needed to assess the main claims. Code and machine-readable experimental records will be released publicly. No external code attachment is required to read the supporting analyses.

AI USE STATEMENT

Generative AI assisted research planning and derivations, coding and synthetic-task implementation, experiment execution and analysis, figure preparation, literature search, and manuscript drafting and editing. The authors take responsibility for the numerical results, references, manuscript, and accompanying artifacts.

REFERENCES

Advani, M. S., Saxe, A. M., and Sompolinsky, H. (2020). High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132, 428–446.

Ahmad, N. (2024). Correlations are ruining your gradient descent. *arXiv:2407.10780*.

Akrout, M., Wilson, C., Humphreys, P., Lillicrap, T., and Tweed, D. (2019). Deep learning without weight transport. In *Advances in Neural Information Processing Systems*.

Amari, S. (1998). Natural gradient works efficiently in learning. *Neural Computation*, 10(2), 251–276.

Amari, S., Ba, J., Grosse, R., Li, X., Nitanda, A., Suzuki, T., Wu, D., and Xu, J. (2021). When does preconditioning help or hurt generalization? In *Proceedings of the International Conference on Learning Representations*.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. (2019). Invariant risk minimization. *arXiv:1907.02893*.

Bartunov, S., Santoro, A., Richards, B. A., Marris, L., Hinton, G. E., and Lillicrap, T. P. (2018). Assessing the scalability of biologically-motivated deep learning algorithms and architectures. In *Advances in Neural Information Processing Systems*, volume 31.

Belkin, M., Hsu, D., and Xu, J. (2020). Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180. doi:10.1137/20M1336072.

Benzing, F. (2022). Gradient descent on neurons and its link to approximate second-order optimization. In *Proceedings of the 39th International Conference on Machine Learning*, PMLR 162, 1817–1853.

Boeshertz, G., Pascanu, R., and Clopath, C. (2026). Overcoming rank collapse in feedback alignment. *arXiv:2606.11123*.

Braun, G., Bao, H., Huang, W., and Imaizumi, M. (2026). Spectral gradient descent mitigates anisotropy-driven misalignment: A case study in phase retrieval. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. arXiv:2601.22652.

Caillon, P., Fagnou, E., Delattre, B., and Allauzen, A. (2026). Scaling direct feedback learning with Jacobian alignment guarantees. In *Proceedings of the International Conference on Learning Representations*.

Chu, D. and Bacho, F. (2023). Random feedback alignment algorithms to train neural networks: Why do they align? *arXiv:2306.02325*.

Dalm, S., Offergeld, J., Ahmad, N., and van Gerven, M. (2024). Efficient deep learning with decorrelated backpropagation. *arXiv:2405.02385*.

Dalm, S., van Gerven, M., and Ahmad, N. (2026). Node perturbation can effectively train multi-layer neural networks. *Neuromorphic Computing and Engineering*, 6, 034020. doi:10.1088/2634-4386/ae9be4.

Desjardins, G., Simonyan, K., Pascanu, R., and Kavukcuoglu, K. (2015). Natural neural networks. In *Advances in Neural Information Processing Systems*, volume 28.

Fiete, I. R., Fee, M. S., and Seung, H. S. (2007). Model of birdsong learning based on gradient estimation by dynamic perturbation of neural conductances. *Journal of Neurophysiology*, 98(4), 2038–2057.

Frenkel, C., Lefebvre, M., and Bol, D. (2021). Learning without feedback: Fixed random learning signals allow for feedforward training of deep neural networks. *Frontiers in Neuroscience*, 15, 629892.

Garg, S. and Vempala, S. (2022). How and when random feedback works: A case study of low-rank matrix factorization. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 4070–4108.

George, T., Laurent, C., Bouthillier, X., Ballas, N., and Vincent, P. (2018). Fast approximate natural gradient descent in a Kronecker-factored eigenbasis. In *Advances in Neural Information Processing Systems*.

Girotti, M., Mitliagkas, I., and Gidel, G. (2021). Convergence analysis and implicit regularization of feedback alignment for deep linear networks. *arXiv:2110.10815*.

Grosse, R. and Martens, J. (2016). A Kronecker-factored approximate Fisher matrix for convolution layers. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 573–582.

Guerguiev, J., Lillicrap, T. P., and Richards, B. A. (2017). Towards deep learning with segregated dendrites. *eLife*, 6, e22901.

Gupta, V., Koren, T., and Singer, Y. (2018). Shampoo: Preconditioned stochastic tensor optimization. In *Proceedings of the International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1842–1850.

Gustafson, K. (2010). On my Min-Max Theorem (1968) and its consequences. *Acta et Commentationes Universitatis Tartuensis de Mathematica*, 14, 45–51.

Heckel, R. and Yilmaz, F. F. (2021). Early stopping in deep networks: Double descent and how to eliminate it. In *Proceedings of the International Conference on Learning Representations*.

Huang, L., Yang, D., Lang, B., and Deng, J. (2018). Decorrelated batch normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 791–800.

Huang, L., Zhou, Y., Zhu, F., Liu, L., and Shao, L. (2019). Iterative normalization: Beyond standardization towards efficient whitening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4874–4883.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456.

Kang, B., Richards, M., and Sabatini, B. (2026). Credit assignment via neural manifold noise correlation. *arXiv:2601.02636*.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*.

Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. Technical report, University of Toronto.

Larochelle, H., Erhan, D., Courville, A., Bergstra, J., and Bengio, Y. (2007). An empirical evaluation of deep architectures on problems with many factors of variation. In *Proceedings of the 24th International Conference on Machine Learning*, pages 473–480.

Launay, J., Poli, I., Boniface, F., and Krzakala, F. (2020). Direct feedback alignment scales to modern deep learning tasks and architectures. In *Advances in Neural Information Processing Systems*, volume 33, pages 9346–9360.

Lee, Y. and Lee, S. (2026). Enabling fine-tuning of direct feedback alignment via feedback-weight matching. In *Proceedings of the International Conference on Learning Representations*.

Lee, D.-H., Zhang, S., Fischer, A., and Bengio, Y. (2015). Difference target propagation. In *Machine Learning and Knowledge Discovery in Databases*.

Liao, Q., Leibo, J., and Poggio, T. (2016). How important is weight symmetry in backpropagation? In *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1), 1837–1844.

Lillicrap, T. P., Cownden, D., Tweed, D. B., and Akerman, C. J. (2016). Random synaptic feedback weights support error backpropagation for deep learning. *Nature Communications*, 7, 13276.

Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J., and Hinton, G. (2020). Backpropagation and the brain. *Nature Reviews Neuroscience*, 21(6), 335–346.

Martens, J. and Grosse, R. (2015). Optimizing neural networks with Kronecker-factored approximate curvature. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2408–2417.

Moskovitz, T. H., Litwin-Kumar, A., and Abbott, L. F. (2018). Feedback alignment in deep convolutional networks. *arXiv:1812.06488*.

Nøkland, A. and Eidnes, L. H. (2019). Training neural networks with local error signals. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4839–4850.

Nøkland, A. (2016). Direct feedback alignment provides learning in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 29.

Refinetti, M., d'Ascoli, S., Ohana, R., and Goldt, S. (2021). Align, then memorise: The dynamics of learning with feedback alignment. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8925–8935.

Roy, A., Apolinario, M. P., Das Biswas, S., and Roy, K. (2026). Feedback alignment meets low-rank manifolds: A structured recipe for local learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2984–2992.

Sacramento, J., Ponte Costa, R., Bengio, Y., and Senn, W. (2018). Dendritic cortical microcircuits approximate the backpropagation algorithm. In *Advances in Neural Information Processing Systems*, volume 31.

Scellier, B. and Bengio, Y. (2017). Equilibrium propagation: Bridging the gap between energy-based models and backpropagation. *Frontiers in Computational Neuroscience*, 11, 24.

Tang, Z., Jiang, F., Gong, M., Li, H., Wu, Y., Yu, F., Wang, Z., and Wang, M. (2021). SKFAC: Training neural networks with faster Kronecker-factored approximate curvature. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13479–13487.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press.

Wang, Z., Müller, K., Filipovich, M., Launay, J., Ohana, R., Pariente, G., Mokaadi, S., Brossollet, C., Moreau, F., Cappelli, A., Poli, I., Carron, I., Daudet, L., Krzakala, F., and Gigan, S. (2026). Streamlined optical training of large-scale modern deep learning architectures with direct feedback alignment. *Proceedings of the National Academy of Sciences*, 123(20), e2532022123.

Whittington, J. C. R. and Bogacz, R. (2017). An approximation of the error backpropagation algorithm in a predictive coding network with local Hebbian synaptic plasticity. *Neural Computation*, 29(5), 1229–1262.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8, 229–256.

Xiao, H., Rasul, K., and Vollgraf, R. (2017). Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv:1708.07747*.

Yoshida, K. and Nitanda, A. (2025). How does preconditioning guide feature learning in deep neural networks? *arXiv:2509.25637*.

SUPPLEMENTARY MATERIAL

The supplement begins with the joint confirmation (Appendix A), including all primary comparisons, selected settings, seed-level values and deviations. The remaining appendices provide the aligned theory and numerical protocols (Appendix B), implementation and historical study provenance (Appendix C), synthetic controls (Appendix D), and short-budget factor studies (Appendix E). Appendices F–J document earlier CIFAR-10 scale, work, damping and stability comparisons. The orientation construction and related mechanisms conclude the supplement (Appendices K–L). Unless a caption specifies sample SD, $\pm$ denotes SEM over the stated replication unit. Each study retains its own analysis family. Earlier studies are not pooled with the joint confirmation. Exploratory architectures that do not support the present claims are omitted.

A JOINT CONFIRMATION: PROTOCOL AND COMPLETE EVIDENCE

The joint confirmation asks which conditioning benefits survive stronger baselines and prospectively fixed comparisons. Development, model selection and scientific interpretation are distinguished from the subsequent test evaluation. This appendix reports the complete primary family, including failed predictions; the historical studies below retain their original protocols and do not enlarge this family retrospectively.

A.1 DESIGN, SELECTION AND REPLICATION

The final training inventory contains 880 models in eight cohorts (Table 2). Each selected setting uses the same ten model seeds, 940101–940110; feedback seeds add 1000. Pairing is within a task and cohort, not across independently selected studies. The real-data splits are fixed, so uncertainty describes training randomness conditional on those splits and choices. Synthetic seeds additionally regenerate the task data. All retained models complete with finite endpoints. The preceding expanded development includes 882 runs: 878 complete and four fail numerically, all at the largest tested FOOF-BP learning rate on CIFAR-10. Failed candidates are ineligible for selection; no selected family is substituted after testing.

Expanded development uses two paired seeds per candidate. Selection minimizes mean best validation CE, then breaks ties by validation accuracy and fixed case order. Work-budget evaluations occur at approximately 5% budget increments, including initialization and the final step. The early-activity CIFAR candidate comes from the preceding matched-work development, selected on validation CE at the same budget. Geometry settings and fixed-epoch comparisons retain their previously frozen validation choices. The constant-rate timing comparison fixes the declared learning rate rather than choosing a favorable rate after inspection. These studies inform one joint confirmation, not independent sequences of test-guided tuning.

For the practical CIFAR factorial, raw SGD searches learning rates $\{0.0003, 0.003, 0.03, 0.3, 3, 30\}$; FOOF crosses rates $\{0.3, 3, 30\}$ with absolute damping $\{0.03, 0.3, 3\}$ for both BP and DFA. AdamW references search $\{10^{-4}, 3\times10^{-4}, 10^{-3}\}$; FD-DFA searches $\{3\times10^{-5}, 10^{-4}, 3\times10^{-4}\}$. A crosses $\{3\times10^{-4}, 10^{-3}, 3\times10^{-3}\}$ with relative damping $\{1/3, 1, 3\}$. Thus BP/DFA factorial grids are balanced, but optimizer families do not have identical numbers of candidates. Selected A and FD settings lie on boundaries; these finite searches do not establish optimal convergence envelopes. The error studies search the common A-selected rate over $\{10^{-4}, 3\times10^{-4}, 10^{-3}\}$ and relative E damping over $\{1, 10, 100\}$, without enlarging an edge selection. Image-benchmark raw AdamW and FD grids use those respective rate sets; A crosses the AdamW grid with $\rho_A \in \{0.3, 3\}$, and FOOF crosses rates $\{0.03, 0.3, 3\}$ with absolute damping $\{0.3, 3\}$.

The common seed list, cases, source and analysis are frozen before confirmation evaluation. All included training finishes and passes the global audit before any joint test evaluation. A separate interpretation map is recorded after development but before test access; it does not change the tests or thresholds and is not represented as a preregistration preceding development. The standard CIFAR-10, MNIST and Fashion-MNIST test sets had historical project exposure. Fresh training seeds do not make those datasets untouched. The original MNIST-background-images test member is first read after the joint gate. No test score selects a setting, horizon, model seed or replacement run.

Cohort	Budget	Models	Test endpoint	GPU
CIFAR practical / FOOF	200 s	100	Best validation CE	H100
Synthetic FOOF	20 s	140	Best validation CE	H200
CIFAR temporary E	200 s	40	Best validation CE	H100
MNIST temporary E	60 s	40	Best validation CE	H200
Image benchmarks	60 s	420	Best validation CE	RTX PRO 6000
CIFAR moment controls	200 epochs	70	Final	H100
CIFAR constant-rate timing	200 epochs	40	Final	H200
CIFAR early-A efficiency	200 epochs	30	Final	H100

Table 2: **Joint confirmation ledger.** Time budgets measure synchronized learning work, not total elapsed experiment time. Every case has ten paired seeds. The final endpoint for the efficiency criterion is specified in the analysis, separately from the validation-selected work-budget endpoint.

A.2 Architecture, operators and work accounting

CIFAR-10 uses a 3072–1024–512–10 ReLU MLP, 45,000 training and 5,000 validation images (split seed 925005), batch size 256, and feedback scale 0.1. Input channel moments are estimated on training images only. Training uses reflected-padding-four random crops and horizontal flips; validation and test do not. Core comparisons share BN before ReLU. FD-DFA uses learned forward decorrelation instead of BN, with decorrelation rate 10^{-5}. AdamW uses $(\beta_1, \beta_2) = (0.9, 0.999)$ and $\epsilon = 10^{-8}$; weight decay is 10^{-4} on weight matrices and zero on biases and BN parameters. The raw SGD/FOOF factorial uses SGD without momentum. Thus the tuned AdamW DFA baseline prevents interpreting an advantage over weak SGD alone as the practical result. The factorial interaction uses each credit rule's SGD reference, not its AdamW reference.

For A, $\lambda_A = \max(\rho_A \operatorname{tr}(C_A)/d, 10^{-6})$, where d is the presynaptic width. Current-minibatch moments condition hidden weight matrices, followed by their own-raw-update norm matching. Biases, readout and BN-affine parameters retain their learner's updates. Centered and mean-preserving controls change only the denominator moment; the raw numerator and uncentered-energy damping convention remain fixed. FOOF instead uses normalized exponential moving moments with decay 0.95, 50 calibration minibatches, and inverse refresh every 100 updates. It conditions every weight matrix, including the readout, without norm matching; bias and BN-affine updates remain SGD. Its absolute damping applies to mean moments. This adapts the published FOOF recipe to the present biased/BN architecture; it is not a reproduction of the original paper's experimental setting.

Work-budget cosine schedules warm up for the first 2.5% of measured work and decay to 1% of the peak rate. Fixed-epoch cosine schedules warm up for five epochs. The timing and temporary-E studies use constant learning rates without warmup or decay. Early and late conditioning occupy the first and last quarters of the declared work or update budget, respectively.

Learning time includes synchronization, sampling and augmentation, FOOF calibration, moment/inverse updates, gradient and optimizer work, and numerical checks. Validation, fixed-state probes, hashing and exports are excluded. Estimated leading-order linear-algebra FLOPs count matrix products, Gram systems, solves, FOOF calibration and inverse refresh, and FD transforms, with fused multiply–add counted as two operations. They exclude pointwise activation, BN reductions, optimizer arithmetic, data transforms and exports. These estimates are not hardware-counter FLOPs, and time matching does not imply arithmetic matching. All training uses float32 with TF32 disabled; evaluation losses use float64 reductions (PyTorch 2.9.1, CUDA 12.8).

A.3 Hardware deviation and interruption handling

Benchmark settings were selected on H200 GPUs. Before any benchmark confirmation run began, queue constraints led us to move the complete 420-model cohort, including all baselines, to RTX PRO 6000 Blackwell GPUs. Selected settings, seeds and the nominal 60-second budget were retained, and the amendment was recorded prospectively. Across 40 eligible case groups with uninterrupted records, confirmation performs about 35% more updates on average than development. Different seeds and case-dependent rates preclude interpreting that ratio as a controlled hardware speedup. Early

checkpoint selection does not remove this change in training horizon. All comparisons within the benchmark confirmation use the same GPU model; the main H100 CIFAR comparison is unaffected.

Five benchmark runs were interrupted and resumed by the scheduler, which violated the uninterrupted-work requirement. Before test access, all five were selected for one cold replay solely from interruption metadata: A on the MNIST mixture at seed 940110, and FOOF-DFA on that mixture at seeds 940106–940109. Original records and checkpoints were archived. Each replay used identical settings, seed, source, data, initialization, feedback and GPU type, and completed in one attempt. The failed audit and replay rule remain in the provenance record; these are replacements for interrupted attempts, not additional independent observations. Fixed-update geometry and timing experiments permit state-preserving resumption and are not used to claim uninterrupted timing efficiency.

A.4 Statistical criteria fixed before testing

The primary family consists of 21 two-sided paired t tests, with Holm correction across the entire family. Table 3 gives individual 95% t intervals and adjusted probabilities; the intervals do not provide simultaneous familywise coverage. A practical CIFAR claim also requires a mean gain of at least one accuracy percentage point and no increase in mean CE. The three practical comparisons pass these criteria. Nonsignificance is not evidence of equivalence.

Early-only efficiency is a separate prespecified noninferiority question at the final fixed-epoch endpoint. It requires a one-sided 95% lower bound for early-minus-full accuracy above -0.5 percentage points and a mean paired work ratio at most 0.8. The observed difference is -0.164 points, the lower bound is -0.401, and the work ratio is 0.658; this criterion passes. The individual two-sided interval is $[-0.456, 0.128]$ points. This is a conditional accuracy–work tradeoff, not equal performance or a schedule-independent law.

Clean and label-noise benchmark predictions use secondary individual 90% paired intervals and a ± 1-point equivalence margin. MNIST clean and both Fashion conditions meet the margin; MNIST label noise does not. These are not a familywise equivalence claim. All 21 primary seed contrasts are given at the end of this appendix, including effects opposite to prediction.

ID	Primary contrast	Mean	Individual 95% CI	p_{Holm}
1	A – tuned DFA	$+2.263$	$[1.958, 2.568]$	6.85×10^{-7}
2	Early A – tuned DFA	$+2.756$	$[2.448, 3.064]$	1.37×10^{-7}
3	FOOF-DFA – tuned DFA	$+3.554$	$[3.240, 3.868]$	1.95×10^{-8}
4	FOOF gain: DFA – BP	-0.229	$[-1.148, 0.690]$	1.000
5	Full – mean-preserving control	$+3.205$	$[3.006, 3.404]$	8.88×10^{-10}
6	Centered – mean-preserving control	$+3.126$	$[2.817, 3.435]$	5.03×10^{-8}
7	Early – late A (DFA)	$+1.186$	$[0.926, 1.446]$	4.07×10^{-5}
8	Early–late effect: DFA – BP	$+0.055$	$[-0.455, 0.565]$	1.000
9	Early K – A: accuracy	$+0.384$	$[0.018, 0.750]$	0.315
10	Early K – A: CE improvement	$+0.0071$	$[0.0019, 0.0122]$	0.155
11	Early K – diagonal E: accuracy	$+0.413$	$[0.025, 0.801]$	0.315
12	Early K – diagonal E: CE improvement	$+0.0086$	$[0.0020, 0.0153]$	0.159
13	Error direction change above 0.001	$+0.3773$	$[0.3749, 0.3796]$	1.04×10^{-18}
14	MNIST mixture: accuracy	-0.755	$[-1.373, -0.137]$	0.198
15	MNIST mixture: CE improvement	-0.0303	$[-0.0473, -0.0134]$	0.038
16	MNIST: nuisance – clean gain	-0.320	$[-0.941, 0.301]$	1.000
17	Fashion mixture: accuracy	$+0.548$	$[0.145, 0.951]$	0.155
18	Fashion mixture: CE improvement	-0.0153	$[-0.0230, -0.0076]$	0.020
19	Fashion: nuisance – clean gain	$+0.467$	$[0.028, 0.906]$	0.315
20	Original backgrounds: accuracy	$+0.143$	$[-0.268, 0.554]$	1.000
21	Original backgrounds: CE improvement	$+0.0033$	$[-0.0049, 0.0115]$	1.000

Table 3: **All 21 primary confirmation contrasts.** Each uses ten paired training seeds. Accuracy effects are percentage points; positive CE effects mean lower loss. The direction row is $1 - \cos(K, A) - 0.001$. Holm correction covers all rows together. Individual intervals are not simultaneous intervals. Negative significant loss effects contradict the positive nuisance prediction.

A.5 Interpretation of the selected comparisons

The primary practical comparisons favor A, early A and FOOF-DFA over tuned DFA, with favorable accuracy differences in all ten seed pairs and lower mean CE. Table 5 includes the SGD factorial, unnormalized baseline and arithmetic costs omitted from the compact main table. The FOOF-by-credit interaction is the difference between the FOOF gain over SGD in DFA and that in BP. Its interval spans zero. The data therefore establish that FOOF works with random feedback, without establishing preferential benefit for DFA. No equivalence test was specified for the FOOF-DFA/FOOF-BP gap.

Rule	Optimizer	η	ρ_A or λ
BP (SGD)	SGD	0.03	–
FOOF-BP	SGD	3	0.3
BP (AdamW)	AdamW	0.0001	–
DFA (SGD)	SGD	0.03	–
FOOF-DFA	SGD	3	0.3
DFA (AdamW)	AdamW	0.0001	–
FD-DFA	AdamW	0.0003	–
Activity DFA	AdamW	0.0003	3
DFA without BN	AdamW	0.0001	–
Early activity DFA	AdamW	0.0005	1

Table 4: **Selected settings for the practical CIFAR-10 cohort.** A uses relative activity damping; FOOF uses absolute damping of mean moments. FD-DFA uses decorrelation rate 10^{-5}. Early A uses conditioning for the first quarter of learning work. These choices precede test evaluation.

Rule	Accuracy (%)	CE	Updates	FLOPs (10^{12})
BP (SGD)	65.70 ± 0.32	0.9999	70,269	283.1
FOOF-BP	70.03 ± 1.17	0.9537	48,830	1476.7
BP (AdamW)	66.88 ± 0.25	0.9631	61,787	249.0
DFA (SGD)	65.11 ± 0.35	1.0108	70,839	266.8
FOOF-DFA	69.21 ± 0.30	0.9540	49,006	1469.2
DFA (AdamW)	65.66 ± 0.25	0.9913	62,035	233.7
FD-DFA	67.01 ± 0.25	0.9602	42,177	437.8
Activity DFA	67.92 ± 0.27	0.9435	35,333	226.2
DFA without BN	61.30 ± 0.60	1.1444	89,829	338.3
Early activity DFA	68.41 ± 0.34	0.9449	55,343	231.8

Table 5: **Complete practical CIFAR-10 cohort.** Accuracy is mean $\pm$ sample SD over ten paired seeds; CE is the mean. Updates and estimated leading-order linear-algebra FLOPs are for the complete training budget, including work after a selected checkpoint; they are not time- or FLOP-equated endpoints.

A.6 Activity statistics

The geometry experiment compares C_A, $\mathrm{Cov}(h)$, $\mathrm{diag}(C_A)$, $\mathrm{diag}(\mathrm{Cov}(h)) + \mu\mu^\top$, and $\mathrm{tr}(\mathrm{Cov}(h))I/d + \mu\mu^\top$, as well as raw DFA and a full-moment sample-space implementation. Full and centered moments both improve over the mean-preserving diagonal-covariance control in every seed. The controls share the raw numerator, norm matching and damping convention, but use separately validation-selected damping. Their comparison supports a role for centered correlations beyond the tested mean-only explanation; it does not show that every diagonal optimizer or mean-based method must fail.

Rule	Accuracy (%)	CE	η	ρ_A or λ
Centered covariance	67.90 ± 0.30	0.9577	0.001	1
Diagonal moment	64.64 ± 0.22	1.0390	0.001	0.1
Diagonal covariance + mean	64.77 ± 0.34	1.0410	0.001	30
Full moment	67.98 ± 0.24	0.9569	0.001	1
Isotropic covariance + mean	64.70 ± 0.32	1.0429	0.001	30
Raw DFA	64.78 ± 0.30	1.0343	0.001	1
Full (sample-space solve)	67.98 ± 0.27	0.9573	0.001	1

Table 6: **Final fixed-epoch geometry endpoints.** Accuracy is mean $\pm$ sample SD over ten paired seeds; CE is the mean. The damping column is relative ρ_A for A/geometry and absolute mean-moment damping λ for FOOF or MNIST A; raw rules have no damping.

A.7 Timing and efficiency

The constant-rate timing interaction compares early-minus-late gains in DFA and BP. The early-window DFA effect survives correction; its extra benefit over BP does not. Shared-state interventions use four development seeds, forks at epochs 50 and 150, and equal 50-epoch continuations with the same future samples and optimizer state. Conditioning can still help after a late fork. These results do not establish irreversible nuisance capture.

Rule	Accuracy (%)	CE	η	ρ_A or λ
DFA, early A	65.76 ± 0.27	1.0023	0.0003	1
DFA, late A	64.58 ± 0.30	1.0543	0.0003	1
BP, early A	67.55 ± 0.34	0.9951	0.0003	1
BP, late A	66.41 ± 0.45	1.0756	0.0003	1

Table 7: **Final constant-rate timing endpoints.** Accuracy is mean $\pm$ sample SD over ten paired seeds; CE is the mean. The damping column is relative ρ_A for A/geometry and absolute mean-moment damping λ for FOOF or MNIST A; raw rules have no damping.

Rule	Accuracy (%)	CE	η	ρ_A or λ
DFA	65.01 ± 0.25	1.0109	0.0003	–
Activity DFA	67.44 ± 0.24	0.9647	0.003	0.3333
Early activity DFA	67.27 ± 0.32	0.9743	0.003	0.3333

Table 8: **Final fixed-epoch efficiency endpoints.** Accuracy is mean $\pm$ sample SD over ten paired seeds; CE is the mean. The damping column is relative ρ_A for A/geometry and absolute mean-moment damping λ for FOOF or MNIST A; raw rules have no damping.

Task	Fork	Prior A	DFA gain (pp)	BP gain (pp)
CIFAR-10	25%	On	2.51 ± 0.92	1.49 ± 0.61
CIFAR-10	25%	Off	1.68 ± 0.85	1.04 ± 0.47
CIFAR-10	75%	On	2.04 ± 0.29	1.10 ± 0.29
CIFAR-10	75%	Off	1.16 ± 0.40	0.71 ± 0.68
Synthetic	25%	On	1.11 ± 0.65	1.82 ± 0.29
Synthetic	25%	Off	1.27 ± 0.31	2.29 ± 0.19
Synthetic	75%	On	0.60 ± 0.18	0.61 ± 0.08
Synthetic	75%	Off	1.04 ± 0.28	0.71 ± 0.28

Table 9: **Shared-checkpoint development interventions.** Mean $\pm$ sample SD of validation-accuracy differences over four paired prefix seeds. At each saved state, conditioning-on and conditioning-off branches receive the same 50-epoch continuation, optimizer state, batches and constant learning rate. These are development results, outside the primary family; late gains contradict an irreversible-capture interpretation.

A.8 TEMPORARY ERROR CONDITIONING

The error test uses the common learning rate selected on A alone and holds it constant. CIFAR uses $\eta = 3 \times 10^{-4}$ and $\rho_A = 1$; early full E selects relative damping $\rho_E = 1$, while diagonal E selects $\rho_E = 100$. MNIST uses a 784–256–128–10 ReLU MLP without BN, batch 128, 55,000/5,000 training/validation examples, feedback scale 1, AdamW without weight decay, $\eta = 10^{-3}$, absolute $\lambda_A = 0.3$, and $\rho_E = 10$ for both E forms. Its images are scaled by $1/255$ without augmentation. Relative E damping scales the mean squared per-example gated local error; minibatch mean-loss scaling is undone before the error moment is formed.

Early K applies E during the first quarter of measured learning work and then continues A. Persistent K uses the selected early-K damping throughout. A positive scalar E factor cancels under hidden-matrix norm matching, so A is also the matched scalar comparator; this identity was numerically checked, not counted as another independent training arm. The directional measure is $1 - \cos(K, A)$ on common states and fixed training batches, averaged over hidden layers and positive-time probes before the early-window boundary, then within each seed. It isolates E rather than conflating E with activity conditioning.

The prespecified E claim requires improved accuracy and CE over both A and diagonal E, and directional change greater than 0.001; all five tests enter the global family. Early K has favorable means over A (0.384 accuracy points and 0.0071 lower CE), but the adjusted probabilities are 0.315 and 0.155. Only the directional criterion passes. The gate therefore fails: the result is a small positive estimate with insufficient corrected evidence, not a negative effect or an exactly zero effect. The secondary MNIST result cannot replace this failed CIFAR criterion. No window, seed or learning-rate extension follows it.

Rule	Accuracy (%)	CE	η	ρ_A or λ
Activity DFA	66.87 ± 0.23	0.9777	0.0003	1
Early K $\rightarrow$ A	67.25 ± 0.39	0.9706	0.0003	1
Early diagonal E $\rightarrow$ A	66.84 ± 0.23	0.9793	0.0003	1
Persistent K	66.47 ± 0.34	0.9703	0.0003	1

Table 10: **Stable-training CIFAR-10 error-factor endpoints.** Accuracy is mean $\pm$ sample SD over ten paired seeds; CE is the mean. The damping column is relative ρ_A for A/geometry and absolute mean-moment damping λ for FOOF or MNIST A; raw rules have no damping.

Rule	Accuracy (%)	CE	η	ρ_A or λ
Activity DFA	97.54 ± 0.11	0.0834	0.001	0.3
Early K $\rightarrow$ A	97.77 ± 0.15	0.0764	0.001	0.3
Early diagonal E $\rightarrow$ A	97.53 ± 0.13	0.0842	0.001	0.3
Persistent K	97.73 ± 0.12	0.0764	0.001	0.3

Table 11: **Secondary MNIST error-factor endpoints.** Accuracy is mean $\pm$ sample SD over ten paired seeds; CE is the mean. The damping column is relative ρ_A for A/geometry and absolute mean-moment damping λ for FOOF or MNIST A; raw rules have no damping.

A.9 IMAGE-BENCHMARK PREDICTIONS

Constructed MNIST and Fashion conditions share 10,000 training and 2,000 validation foreground identities (training-only split seed 934005). The clean condition scales pixels by $1/255$. Label noise changes 20% of training labels independently to a different class; validation and test labels remain clean. Input nuisance mixes 0.35 foreground with 0.65 independently assigned grayscale CIFAR-10 training-image background, center-cropped to 28×28. The background pools for training, validation and test are disjoint and assignment does not use foreground class. These are paired constructed variants, not the original MNIST-variations dataset. The original MNIST-background-images archive (Larochelle et al., 2007) supplies a separate standard task: its first 10,000 training rows train, its remaining 2,000 rows validate, and its 50,000-row test member is evaluated after the global gate.

All image-benchmark conditions use 784–256–128–10 ReLU MLPs, batch 128, feedback scale 1, no weight decay and no further augmentation. Core arms use BN; FD uses decorrelation. Before testing we predict accuracy and loss gains from A under input nuisance, a larger accuracy gain than under clean inputs, and small clean/label-noise effects. Neither constructed dataset meets the full prediction. CE worsens significantly after correction in both, and the original background task provides no corrected benefit. The complete six-method endpoints below retain these reversals. The task description alone does not establish the spectral relation assumed by the theory.

Rule	Accuracy (%)	CE	η	Damp.	Updates	TFLOPs
MNIST clean						
DFA	94.80 ± 0.18	0.1732	0.001	–	30,786	3.69
BP	95.83 ± 0.28	0.1465	0.001	–	30,501	3.89
FD-DFA	94.83 ± 0.24	0.1907	0.0003	–	23,146	7.68
Activity DFA	94.37 ± 0.37	0.1889	0.001	3	17,248	3.93
FOOF-BP	95.96 ± 0.23	0.1567	3	3	27,915	17.90
FOOF-DFA	95.33 ± 0.14	0.1721	0.3	3	28,084	17.79
MNIST label noise						
DFA	90.48 ± 0.28	0.4837	0.0001	–	29,180	3.50
BP	90.02 ± 0.71	0.4864	0.0003	–	30,700	3.91
FD-DFA	91.06 ± 0.89	0.4354	0.0003	–	21,290	7.07
Activity DFA	89.41 ± 0.42	0.4966	0.0003	3	18,295	4.17
FOOF-BP	90.50 ± 0.21	0.4837	0.03	3	28,007	17.96
FOOF-DFA	79.02 ± 3.17	0.9634	3	3	28,056	17.77
MNIST mixture						
DFA	79.07 ± 0.35	0.6731	0.0001	–	30,802	3.69
BP	79.18 ± 0.37	0.6638	0.0001	–	30,968	3.95
FD-DFA	76.95 ± 1.17	0.7892	0.0001	–	22,931	7.61
Activity DFA	78.31 ± 0.83	0.7035	0.001	3	18,529	4.22
FOOF-BP	81.49 ± 0.47	0.7268	0.3	0.3	28,091	18.01
FOOF-DFA	79.01 ± 0.32	0.7072	0.03	0.3	28,160	17.84

Table 12: **Image-benchmark confirmation.** All six prespecified families and conditions are retained. Accuracy is mean $\pm$ sample SD; CE, update counts and estimated linear-algebra work are means over ten seeds. All runs use the same 60-second RTX PRO 6000 budget and validation-CE selection. Damping is relative for A and absolute for FOOF. FD uses decorrelation rate 10^{-5}.

Rule	Accuracy (%)	CE	η	Damp.	Updates	TFLOPs
Fashion clean						
DFA	84.91 ± 0.25	0.4466	0.0001	–	30,857	3.70
BP	85.67 ± 0.27	0.4263	0.0001	–	30,999	3.95
FD-DFA	82.39 ± 1.59	0.7129	0.0003	–	23,178	7.70
Activity DFA	84.99 ± 0.25	0.4534	0.001	3	18,556	4.23
FOOF-BP	85.08 ± 0.19	0.4356	0.03	3	28,107	18.02
FOOF-DFA	84.44 ± 0.41	0.4951	0.03	0.3	26,025	16.49
Fashion label noise						
DFA	81.73 ± 0.39	0.6872	0.0001	–	28,558	3.42
BP	82.50 ± 0.75	0.6704	0.0001	–	30,285	3.86
FD-DFA	76.16 ± 1.40	0.8341	0.0003	–	23,130	7.68
Activity DFA	81.85 ± 0.21	0.6797	0.0001	0.3	18,438	4.20
FOOF-BP	82.17 ± 0.37	0.6733	0.03	3	27,789	17.82
FOOF-DFA	75.82 ± 2.91	0.8441	0.03	0.3	27,626	17.50
Fashion mixture						
DFA	71.47 ± 0.41	0.8266	0.0001	–	30,695	3.68
BP	70.21 ± 0.83	0.8660	0.0001	–	30,534	3.89
FD-DFA	65.50 ± 0.91	1.0977	0.0003	–	23,022	7.64
Activity DFA	72.02 ± 0.43	0.8419	0.0003	3	18,564	4.23
FOOF-BP	71.30 ± 0.29	0.8685	0.03	0.3	27,884	17.88
FOOF-DFA	70.87 ± 0.75	0.9376	0.03	0.3	28,146	17.83

Table 13: **Image-benchmark confirmation (continued).** All six prespecified families and conditions are retained. Accuracy is mean $\pm$ sample SD; CE, update counts and estimated linear-algebra work are means over ten seeds. All runs use the same 60-second RTX PRO 6000 budget and validation-CE selection. Damping is relative for A and absolute for FOOF. FD uses decorrelation rate 10^{-5}.

Rule	Accuracy (%)	CE	η	Damp.	Updates	TFLOPs
Original MNIST backgrounds						
DFA	73.85 ± 0.58	0.8137	0.0001	–	30,583	3.66
BP	72.95 ± 0.50	0.8462	0.0001	–	31,044	3.96
FD-DFA	71.20 ± 1.22	1.0307	0.0003	–	22,788	7.57
Activity DFA	73.99 ± 0.14	0.8103	0.0001	3	18,548	4.23
FOOF-BP	72.80 ± 0.19	0.8506	0.03	3	28,084	18.00
FOOF-DFA	71.49 ± 0.63	0.9641	0.03	0.3	27,340	17.32

Table 14: **Image-benchmark confirmation (continued).** All six prespecified families and conditions are retained. Accuracy is mean $\pm$ sample SD; CE, update counts and estimated linear-algebra work are means over ten seeds. All runs use the same 60-second RTX PRO 6000 budget and validation-CE selection. Damping is relative for A and absolute for FOOF. FD uses decorrelation rate 10^{-5}.

A.10 SECONDARY SYNTHETIC FACTORIAL

The secondary synthetic FOOF interaction is larger without BN (11.04 points, individual 95% interval $[9.84, 12.24]$) than with BN (3.84 points, $[-0.06, 7.73]$); it is outside the primary family and does not establish a general mechanism specific to DFA.

Rule	Accuracy (%)	CE	η	ρ_A or λ
bp sgd none	15.07 ± 0.56	2.5119	0.001	–
foof bp none	23.37 ± 1.75	2.3998	0.003	0.3
bp tuned none	15.24 ± 0.90	2.5493	$3e-05$	–
dfa sgd none	16.42 ± 1.21	2.8564	0.001	–
foof dfa none	35.75 ± 1.89	2.2066	0.03	0.3
dfa tuned none	15.58 ± 1.00	2.8502	$3e-05$	–
bp sgd bn	14.26 ± 0.48	2.2351	0.001	–
foof bp bn	19.67 ± 1.72	2.1597	0.003	0.3
bp tuned bn	15.19 ± 0.93	2.2401	$3e-05$	–
dfa sgd bn	19.70 ± 4.25	2.2323	0.001	–
foof dfa bn	28.94 ± 3.07	2.1097	0.003	0.3
dfa tuned bn	15.61 ± 1.81	2.2513	0.0001	–
activity none	34.39 ± 2.02	2.0405	0.0001	0.03
Activity DFA	41.72 ± 2.13	1.6317	0.0003333	0.3

Table 15: **Secondary synthetic FOOF factorial.** Accuracy is mean $\pm$ sample SD over ten paired seeds; CE is the mean. The damping column is relative ρ_A for A/geometry and absolute mean-moment damping λ for FOOF or MNIST A; raw rules have no damping.

A.11 SEED-LEVEL PRIMARY CONTRASTS

Tables 16–17 list every paired value, with IDs matching Table 3. Display rounding does not enter the statistical calculations.

ID	940101	940102	940103	940104	940105
1	$+1.770$	$+2.640$	$+2.120$	$+2.380$	$+3.130$
2	$+2.600$	$+3.170$	$+2.460$	$+3.370$	$+3.290$
3	$+2.910$	$+3.910$	$+4.170$	$+3.840$	$+3.510$
4	$+0.820$	$+1.100$	-0.610	-0.440	-1.370
5	$+3.380$	$+3.470$	$+2.950$	$+3.080$	$+3.230$
6	$+3.010$	$+2.880$	$+2.740$	$+3.590$	$+3.480$
7	$+0.990$	$+1.480$	$+0.980$	$+1.070$	$+1.990$
8	-0.200	-0.380	-1.300	$+0.300$	$+0.780$
9	$+0.270$	-0.330	$+0.460$	$+0.560$	-0.290
10	-0.0008	$+0.0017$	$+0.0165$	$+0.0135$	-0.0018
11	$+0.280$	-0.330	$+0.370$	$+0.380$	$+0.240$
12	-0.0026	$+0.0049$	$+0.0181$	$+0.0084$	$+0.0070$
13	$+0.3771$	$+0.3770$	$+0.3774$	$+0.3723$	$+0.3795$
14	-0.700	-2.940	-0.710	$+0.220$	-0.520
15	-0.0238	-0.0921	-0.0381	-0.0068	-0.0277
16	-0.450	-2.060	-0.450	$+0.430$	$+0.210$
17	$+0.470$	$+1.120$	$+0.780$	-0.030	$+1.020$
18	-0.0089	-0.0033	-0.0106	-0.0345	-0.0171
19	$+0.760$	$+1.050$	$+1.320$	$+0.350$	$+0.760$
20	$+0.108$	$+0.364$	-0.294	-0.344	$+0.216$
21	-0.0036	$+0.0062$	-0.0096	-0.0074	$+0.0062$

Table 16: **Seed-level primary contrasts.** IDs and units match Table 3. These are training replications on fixed real-data splits, not independent datasets.

ID	940106	940107	940108	940109	940110
1	+2.300	+1.940	+1.710	+2.160	+2.480
2	+2.920	+2.030	+2.560	+2.760	+2.400
3	+3.930	+3.650	+3.030	+3.590	+3.000
4	−2.680	+0.150	+1.400	+0.580	−1.240
5	+3.270	+3.760	+3.040	+2.810	+3.060
6	+3.290	+3.760	+3.210	+2.990	+2.310
7	+1.140	+1.080	+1.470	+0.900	+0.760
8	+0.130	−0.480	+1.010	−0.170	+0.860
9	+0.730	+0.490	+1.390	−0.020	+0.580
10	+0.0119	+0.0148	+0.0072	+0.0097	−0.0019
11	+0.700	−0.150	+1.610	+0.200	+0.830
12	+0.0172	+0.0068	+0.0258	+0.0040	−0.0032
13	+0.3755	+0.3755	+0.3829	+0.3817	+0.3738
14	−0.810	−0.980	+0.150	−0.510	−0.750
15	−0.0238	−0.0331	−0.0117	−0.0170	−0.0292
16	−1.050	−0.380	+0.860	+0.510	−0.820
17	−0.420	−0.170	+0.840	+1.130	+0.740
18	−0.0323	−0.0167	−0.0037	−0.0158	−0.0101
19	−0.580	−0.460	+0.340	+0.830	+0.300
20	−0.760	+1.248	+0.736	+0.294	−0.138
21	−0.0114	+0.0223	+0.0183	+0.0092	+0.0032

Table 17: **Seed-level primary contrasts (continued).** IDs and units match Table 3. These are training replications on fixed real-data splits, not independent datasets.

B LINEAR ANALYSIS AND SIMULATION PROTOCOLS

Write $C_{\ell-1} := C_{A,\ell}$ for the activity second moment. To make its role in Eq. 3 precise, consider the minimal setting in which BP, DFA, and nDFA differ. Let the input covariance be positive definite and the inputs satisfy $x \sim \mathcal{N}(0, \Sigma)$, let the target be a linear map $y = W^\star x$, and consider a single hidden layer

$$h = W^{(1)}x, \qquad \hat{y} = W^{(2)}h, \qquad \mathcal{L} = \tfrac{1}{2}\,\mathbb{E}_x\|\hat{y} - y\|^2. \tag{5}$$

Write the residual operator $\varepsilon := W^{(2)}W^{(1)} - W^\star$ and the per-sample output error $e(x) = \varepsilon x$. Let $B \in \mathbb{R}^{n_h \times n_y}$ be the fixed feedback matrix. The expected per-step update directions in $W^{(1)}$ are

$$g^{\mathrm{BP}} = -\,\mathbb{E}_x\big[(W^{(2)})^\top e(x)\, x^\top\big] = -(W^{(2)})^\top \varepsilon\, \Sigma, \tag{6}$$

$$g^{\mathrm{DFA}} = -\,\mathbb{E}_x\big[B\, e(x)\, x^\top\big] = -B\, \varepsilon\, \Sigma, \tag{7}$$

$$g^{\mathrm{nDFA}} = g^{\mathrm{DFA}}\,(\Sigma + \lambda_A I)^{-1} = -B\, \varepsilon\, \Sigma\,(\Sigma + \lambda_A I)^{-1}. \tag{8}$$

The updates differ in (a) which left projection of ε is used ($(W^{(2)})^\top$ for BP, B for DFA/nDFA), and (b) which input-statistics weight is applied on the right (Σ for BP and DFA, and $\Sigma(\Sigma + \lambda_A I)^{-1}$ for nDFA). As $\lambda_A \to 0$ the right factor of nDFA approaches the identity. In this population linear model, conditioning removes the input-statistics weighting Σ that DFA inherits from BP while leaving the feedback direction untouched.

The aligned regime. At initialization, when the forward weights are independent of the zero-mean feedback draw, Eq. 7 gives $\mathbb{E}_B[g^{\mathrm{DFA}}] = 0$. Useful hidden updates can emerge as *feedback alignment* makes the forward weights correlate with B, so that the effective error projection B acquires a positive component along $(W^{(2)})^\top$ (Lillicrap et al., 2016; Refinetti et al., 2021). We state the spectral proposition under the explicit perfect-alignment condition $B = \alpha(W^{(2)})^\top$ with $\alpha > 0$, where $g^{\mathrm{DFA}} \propto g^{\mathrm{BP}}$. Scalar output alone does not imply this alignment. We hold the readout fixed when deriving the training dynamics. More general multiclass alignment can be decomposed into an aligned component plus residual feedback components; the same input-side preconditioner acts on all components, but descent then depends on the aligned component being large enough.

What conditioning preconditions. The loss is quadratic in $W^{(1)}$ with Hessian (in row-major $\mathrm{vec}(W^{(1)})$ coordinates, so the activity factor appears on the right of the Kronecker product)

$$H \;=\; (W^{(2)\top} W^{(2)}) \otimes \Sigma, \tag{9}$$

so a Kronecker-factored Newton step contains an input-side factor Σ^{-1}. Right-multiplication by $(\Sigma + \lambda_A I)^{-1}$ is a damped approximation to that one factor; nDFA does not estimate or invert the output-side Hessian factor $W^{(2)\top} W^{(2)}$. The same input factor appears in a Kronecker-factored natural-gradient construction only under an explicitly specified likelihood (for example homoscedastic Gaussian regression) and the usual factorization approximation (Amari, 1998; Martens and Grosse, 2015). For the nonlinear softmax classifiers below, nDFA is therefore best described as a one-sided, curvature-motivated inverse-second-moment preconditioner, not as an exact Newton or natural-gradient method.

Proposition 1 (input-factor spectral identity). Assume the explicitly aligned linear-Gaussian regime above with fixed readout, population second moments, fixed feedback $B = \alpha(W^{(2)})^\top$ with $\alpha > 0$, and damping $\lambda_A > 0$. Then input-side conditioning replaces the per-eigendirection factor λ_i of raw BP/DFA by $\lambda_i/(\lambda_i + \lambda_A)$, reducing the input-side condition number from $\kappa(\Sigma)$ to κ_{nDFA} in Eq. 10. The condition-number ratio is $\rho(\Sigma, \lambda_A) = \kappa(\Sigma)/\kappa_{\mathrm{nDFA}}$. This is an exact spectral identity for the stated population model. It is not an iteration-complexity result for DFA: realized improvement additionally depends on the residual feedback component, initialization, step size, stopping rule, nonlinear gates, and the jointly estimated minibatch gradient and second moment.

Anisotropy controls the gain. The consequence is a statement about the *conditioning of the training dynamics*, not about the single-step inner product with g^{BP}. Decomposing $\Sigma = U\Lambda U^\top$ with $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$, the per-direction gain of the aligned BP/DFA update on coordinate i scales as λ_i, so the input-side condition number of the update operator is $\kappa(\Sigma) = \lambda_{\max}/\lambda_{\min}$. nDFA replaces this gain by $\lambda_i/(\lambda_i + \lambda_A)$, giving

$$\kappa_{\mathrm{nDFA}} \;=\; \frac{\max_i \lambda_i/(\lambda_i + \lambda_A)}{\min_i \lambda_i/(\lambda_i + \lambda_A)} \;\xrightarrow[\lambda_A \to 0]{}\; 1, \qquad \rho(\Sigma, \lambda_A) := \frac{\kappa(\Sigma)}{\kappa_{\mathrm{nDFA}}} \tag{10}$$

the condition-number reduction factor. Since standard first-order GD bounds on a quadratic scale as $O(\kappa \log \frac{1}{\epsilon})$, $\rho(\Sigma, \lambda_A)$ is the ratio of the corresponding worst-case condition-number bounds. It therefore summarizes the maximum improvement suggested by that analysis, not an exact upper bound on the observed iteration ratio. Whether this spectral improvement translates into faster training depends on *where the task lies* in Σ's spectrum:

(i) Isotropy. If $\Sigma = \sigma^2 I$ then $\rho = 1$: conditioning provides no advantage because every direction already has equal curvature.

(ii) A scoped empirical hypothesis. Preconditioning equalizes the per-direction gain, so it accelerates a task direction in proportion to how *slow* the unconditioned update was there—i.e., to how small that direction's variance λ_i is. Decompose the target into Σ's eigenbasis. If the task lies in the *low-variance* directions of Σ—equivalently, the high-variance directions are task-irrelevant nuisance—the unconditioned update crawls along the task directions and nDFA can realize a speed-up close to the condition-number ratio ρ. If instead the task lies in the *high-variance* directions, the unconditioned update is already fast there and nDFA gives little benefit. Holding $\kappa(\Sigma)$ fixed and only moving the target's energy between low- and high-variance eigendirections changes the realized nDFA speed-up by more than an order of magnitude (Fig. 1D, at fixed $\kappa = 50$). This idealized calculation motivates the design principle behind Eq. 3: noise and nuisance help conditioning not by raising κ per se, but by loading the high-variance directions with task-irrelevant variance, which pushes the task into the low-variance subspace where preconditioning pays off. $\kappa(\Sigma)$ bounds the spectral improvement factor; it is not a general bound on the observed training-speed ratio.

(iii) Excessive damping. As $\lambda_A \to \infty$ the input-statistics weight $\Sigma(\Sigma + \lambda_A I)^{-1} \to \lambda_A^{-1}\Sigma$, hence $g^{\mathrm{nDFA}} \to \lambda_A^{-1} g^{\mathrm{DFA}}$ and $\rho \to 1$: heavy damping reverts nDFA to a scaled DFA step. In population, smaller damping removes the input factor more completely. In the implemented minibatch algorithm, damping can still be required because the empirical factor may be singular or poorly conditioned and because label noise, nonlinear gates, nonstationarity, and finite precision perturb the idealized model. We therefore treat damping as an empirical rule parameter rather than claiming a universal optimum; §B.1 makes this scope explicit.

Population simulation protocol. Figure 1D compares task orientations at fixed $\kappa = 50$. Inputs have dimension 12, the hidden layer has 10 units, and the output has dimension 4. The population moment has eigenvalues spaced geometrically from 1 to 50 and a random orthogonal eigenbasis. A fixed readout has independent Gaussian entries with variance $1/10$, and $B = W^{(2)\top}$. The target is a Gaussian linear map on either the four lowest-variance or four highest-variance eigenvectors, scaled by $1/\sqrt{12}$. Initial first-layer weights have independent Gaussian entries with standard deviation 0.01; the initial residual can therefore occupy other modes too. Activity damping is 10^{-3}, and loss is the population quantity $L = \mathrm{tr}(\varepsilon\Sigma\varepsilon^\top)/2$.

For each rule and initialization we search 24 learning rates spaced geometrically from 10^{-4} to 5, reporting the fewest updates to $L/L_0 < 10^{-6}$. Each candidate has a 4,000-update limit and stops if loss is nonfinite or exceeds $10^3 L_0$. The figure averages seeds 0, 1 and 2. This separately rate-selected illustration shows dependence on task orientation; it is not a matched-computation comparison or a condition-number sweep. The generator will be released with the reproduction code.

Interpreting the projected step. The projected BP-step ratio $\Pi(g, g^{\mathrm{BP}}) = \langle g, g^{\mathrm{BP}}\rangle / \|g^{\mathrm{BP}}\|^2$ measures instantaneous descent in BP units. At fixed norm it is maximized by the BP direction, and it also changes under scalar rescaling of an update. Preconditioning can lower this projection while improving convergence through more balanced learning rates. We therefore interpret Π alongside update norms and trajectories, rather than as an objective to maximize.

Scope and what this does not show. This is a linearized statement about the post-alignment regime with a Gaussian-input model and population covariances. It does not bound the full nonlinear trajectory, model forward-weight alignment to fixed B, or analyze convolutional weight sharing; finite-sample $\widehat{\Sigma}$ estimation and damping are discussed in §B.1 below. It also equates the Fisher with the Hessian only because the model is linear-Gaussian; for the softmax/cross-entropy classifiers used in our experiments, $C_{\ell-1}$ is no longer an exact Fisher factor, so conditioning is a curvature-motivated heuristic there rather than an exact natural gradient. It does give a falsifiable prediction: the improvement over raw DFA at a fixed training budget should grow with the fraction of the layer's second-moment energy that lies in task-irrelevant (nuisance) directions. The bound $\rho \le \kappa(\Sigma)$ concerns the population input-factor condition number, not realized nonlinear speed or accuracy, while any advantage over a tuned BP baseline should disappear under isotropy or when the high-variance directions are themselves task-relevant. The synthetic regimes that gain most over raw DFA from conditioning—nuisance-dominant, low-sample/noisy, mixed-context—are precisely those constructed so that the high-variance hidden directions carry nuisance, while the task-aligned control places the task in the high-variance directions and removes the BP advantage (a slight loss to BP), even though both have anisotropic hidden statistics (Appendix D.1).

B.1 Finite-sample damping: scope of the implemented estimator

For a minibatch with activation matrix X and residual matrix E, the implemented linear-model update uses the *same samples* in its raw outer-product gradient and its activity factor. In the aligned noiseless case,

$$\widehat{G} = -B\varepsilon\widehat{C}, \qquad \widehat{G}(\widehat{C} + \lambda_A I)^{-1} = -B\varepsilon\widehat{C}(\widehat{C} + \lambda_A I)^{-1}, \qquad \widehat{C} = X^\top X/n.$$

Consequently, at $\lambda_A = 0$ and full-rank $\widehat{C}$, the empirical factor cancels exactly. The estimator is not $\Sigma(\widehat{C} + \lambda_A I)^{-1}$, which would combine a population gradient numerator with an independently sampled denominator. Covariance sampling error by itself therefore does not imply inverse blow-up or a universal interior optimum for the batch algorithm.

Damping nevertheless remains necessary when $\widehat{C}$ is singular or ill-conditioned and can be useful under label or gradient noise, nonlinear gating, changing representations, finite precision, and stochastic optimization (Vershynin, 2018). These effects depend jointly on the gradient and factor estimates, so we select λ_A empirically and do not claim a closed-form optimal damping law. Risk and stability bounds for this joint estimator remain open. In particular, this paper draws no depth-, width-, or ImageNet-scaling conclusion from a covariance- concentration formula.

B.2 MODE TIMING: GENERALIZATION DYNAMICS UNDER CONDITIONING

The spectral identity above concerns optimization geometry. This subsection gives a separate, illustrative calculation of test-error trajectories in the same aligned linear-Gaussian model with label noise. It is not a model of the pre-alignment DFA trajectory or evidence for the nonlinear-network results. Within this restricted model, conditioning changes when each eigendirection is learned without changing the least-squares destination.

Finite-sample setup. Keep the model of Eqs. 6–8 with scalar output and explicit alignment $B = \alpha(W^{(2)})^\top$, $\alpha > 0$, and add label noise and a finite training set: n samples $x_\mu \sim \mathcal{N}(0, \Sigma)$, labels $y_\mu = w^{\star\top} x_\mu + \xi_\mu$ with $\xi_\mu \sim \mathcal{N}(0, \sigma^2)$ i.i.d. Collect $X \in \mathbb{R}^{n\times d}$, $\widehat{\Sigma} = X^\top X/n$, and let the first layer follow gradient flow on the *empirical* loss with $W^{(2)}$ fixed on the timescale analyzed (the post-alignment regime; the acquisition of forward-weight alignment to fixed B is outside scope, as in Proposition 1). The composite predictor $w := (W^{(2)}W^{(1)})^\top$ then obeys

$$\dot{w} \;=\; -\,\eta_{\text{eff}}\, M\Big[\widehat{\Sigma}\,(w - w^\star) - \tfrac{1}{n} X^\top \xi\Big], \qquad M = \begin{cases} I & \text{BP, aligned DFA,} \\ (\widehat{\Sigma} + \lambda_A I)^{-1} & \text{nDFA,} \end{cases} \tag{11}$$

with $\eta_{\text{eff}} = \eta\|W^{(2)}\|^2$ for BP and $\eta\alpha\|W^{(2)}\|^2$ for aligned DFA: *BP and aligned DFA share the same clock up to a scalar*. More generally, a scalar-output composite predictor obeys this flow for arbitrary fixed B when $W^{(2)}B > 0$, with $\eta_{\text{eff}} = \eta W^{(2)}B$; this weaker condition does not make the first-layer DFA update proportional to BP. (M and $\widehat{\Sigma}$ commute, so Eq. 11 is the composite form of right-multiplying the first-layer update by the damped inverse, Eq. 8, here with the plug-in $\widehat{\Sigma}$; it is held fixed here to isolate timing.) For $n \geq d$, $\widehat{\Sigma} \succ 0$ a.s. and every rule has the *same* unique fixed point, the least-squares estimator

$$\widehat{w} \;=\; w^\star + \widehat{\Sigma}^{-1}\tfrac{1}{n} X^\top \xi, \qquad \operatorname{Cov}\big(\widehat{w} - w^\star \mid X\big) = \tfrac{\sigma^2}{n}\,\widehat{\Sigma}^{-1}. \tag{12}$$

In the eigenbasis of $\widehat{\Sigma}$ (eigenvalues $\hat{\lambda}_i$), with $w(0) = 0$,

$$w_i(t) \;=\; \widehat{w}_i\,\big(1 - e^{-r_i t}\big), \qquad r_i = \begin{cases} \eta_{\text{eff}}\,\hat{\lambda}_i & \text{BP/DFA,} \\ \eta_{\text{eff}}\,\dfrac{\hat{\lambda}_i}{\hat{\lambda}_i + \lambda_A} & \text{nDFA.} \end{cases} \tag{13}$$

(Units: raw η_{eff} carries $[\text{variance}\cdot\text{time}]^{-1}$, nDFA's carries $[\text{time}]^{-1}$; the minimum risk along each trajectory and within-rule mode-timing ratios are invariant under global time rescaling. Absolute flow times use the stated effective learning rate and are not computational-cost comparisons.)

Exact empirical risk and the population-spectrum surrogate. Define the excess test risk $R(t) = \|w(t) - w^\star\|_\Sigma^2$ and $A_t = \exp(-\eta_{\text{eff}} M\widehat{\Sigma} t)$. Conditional on the design,

$$\begin{aligned} \mathbb{E}_\xi[R(t) \mid X] &= w^{\star\top} A_t^\top \Sigma A_t w^\star \\ &\quad + \frac{\sigma^2}{n}\operatorname{tr}\Big[\Sigma(I - A_t)\widehat{\Sigma}^{-1}(I - A_t)^\top\Big]. \end{aligned} \tag{14}$$

The shared endpoint is $(\sigma^2/n)\operatorname{tr}(\Sigma\widehat{\Sigma}^{-1})$. Averaging over Gaussian designs gives $\sigma^2 d/(n - d - 1)$ for $n > d + 1$ (Belkin et al., 2020). Thus $n \gtrsim d$ alone does not justify replacing inverse empirical moments by population moments. The simpler calculation below replaces the empirical eigensystem by that of Σ; it is a population-spectrum surrogate, accurate only when the relevant sample moments are sufficiently concentrated (Advani et al., 2020). In this surrogate,

$$R_{\text{pop}}(t) \;=\; \underbrace{\sum_i \lambda_i w_i^{\star 2}\, e^{-2r_i t}}_{\text{signal fitting (decreasing)}} \;+\; \underbrace{\frac{\sigma^2}{n}\sum_i \big(1 - e^{-r_i t}\big)^2}_{\text{noise fitting (increasing)}}. \tag{15}$$

Here the rates use population eigenvalues. The empirical trajectories in the numerical comparison instead use Eq. 14. Two readings of Eq. 15 motivate the two-block calculation. First, within this surrogate, every fitted mode eventually contributes σ^2/n of excess risk, independent of its eigenvalue. Second, the exact fixed-point identity shows that all rules share a destination, so conditioning changes the path rather than the least-squares estimator.

Proposition 2 (mode timing in the aligned linear model). Assume the aligned regime above, $n \geq d$, gradient flow on the empirical loss from $w(0) = 0$, Gaussian label noise σ^2, and the population-spectrum approximation of Eq. 15. Define the time for mode i to reach a fraction $0 < \theta < 1$ of its destination, $t_i(\theta) = \ln\big(\frac{1}{1-\theta}\big)/r_i$. Then:

(i) Timing. BP and aligned DFA fit mode i at rate $r_i \propto \lambda_i$ (high-variance modes first: spectral bias), with $t_i/t_j = \lambda_j/\lambda_i$; nDFA fits at rate $r_i \propto \lambda_i/(\lambda_i + \lambda_A)$, with

$$\frac{t_i}{t_j} = \frac{\lambda_j\,(\lambda_i + \lambda_A)}{\lambda_i\,(\lambda_j + \lambda_A)} \;\in\; \Big(1,\; \frac{\lambda_j}{\lambda_i}\Big) \quad \text{for } \lambda_j > \lambda_i, \tag{16}$$

monotonically compressed toward 1 as $\lambda_A \downarrow 0$: the fitting order is preserved but the timing gaps collapse.

(ii) Destinations. All three rules converge to the same $\widehat{w}$ and hence the same final risk (given by Eq. 14; $\sigma^2 d/n$ in the surrogate): conditioning re-times learning without changing what is learned.

(iii) Trajectory optimum and crossing. Take the two-block spectrum: d_T task modes at λ_T carrying all the signal $S := \sum_{i\in T} \lambda_i w_i^{\star 2}$, and d_N noise-only nuisance modes at λ_N. Write $N := \sigma^2 d_N/n$, $N_T := \sigma^2 d_T/n$, $N_{\mathrm{tot}} := N + N_T$, task residual $u := e^{-r_T t}$, and timing ratio $\rho := r_N/r_T$ (not the condition-number ratio of Proposition 1; within this subsection ρ always denotes the timing ratio), so that

$$\mathbb{E}_\xi R = Su^2 + N_T(1-u)^2 + N\big(1-u^\rho\big)^2, \qquad \rho_{\mathrm{DFA}} = \frac{\lambda_N}{\lambda_T}, \quad \rho_{\mathrm{nDFA}}(\lambda_A) = \frac{\lambda_N(\lambda_T + \lambda_A)}{\lambda_T(\lambda_N + \lambda_A)}. \tag{17}$$

The best-achievable risk along the trajectory, $R^\dagger(\rho) := \min_{t\geq 0} \mathbb{E}_\xi R(t)$, is *strictly increasing in* ρ whenever $\sigma > 0$ and $S > 0$. Consequently, for $\sigma, S > 0$ and every finite $\lambda_A > 0$, conditioning strictly lowers $R^\dagger$ iff the task modes are the slow ones ($\lambda_T < \lambda_N$), strictly raises it iff $\lambda_T > \lambda_N$, and leaves it unchanged iff $\lambda_T = \lambda_N$ (it is also unchanged in the degenerate cases $\sigma = 0$ or $S = 0$). In the strong-separation/strong-compression limit ($\rho_{\mathrm{DFA}} \to \infty$, $\rho_{\mathrm{nDFA}} \to 1$), which is ensured by $\lambda_N/\lambda_T \to \infty$ and $\lambda_A/\lambda_T \to 0$,

$$R^\dagger_{\mathrm{DFA}} = \min\Big\{S,\; N + \tfrac{SN_T}{S+N_T}\Big\}, \qquad R^\dagger_{\mathrm{nDFA}} = \frac{S\,N_{\mathrm{tot}}}{S + N_{\mathrm{tot}}}, \qquad \Delta \approx \frac{\min(S,N)^2}{S+N} \;\; (N_T \ll N), \tag{18}$$

To leading order when $N_T \ll N$, the relative improvement $\Delta/R^\dagger_{\mathrm{DFA}} \leq \frac{1}{2}$ is maximized when the signal power matches the accumulated noise floor, $S = \sigma^2 d_N/n$. At finite damping, define $f_\rho(u) = Su^2 + N_T(1-u)^2 + N(1-u^\rho)^2$. The total risk improvement in the two-block surrogate is

$$\Delta(\lambda_A) = \min_{0\leq u\leq 1} f_{\rho_{\mathrm{DFA}}}(u) - \min_{0\leq u\leq 1} f_{\rho_{\mathrm{nDFA}}(\lambda_A)}(u). \tag{19}$$

It includes signal bias, task-noise fitting and nuisance-noise fitting. We evaluate the two complete minima numerically. Their difference depends on signal and noise power as well as the rate ratio; the model does not supply a universal task-eigenvalue threshold for useful damping.

Proof sketch. (i)–(ii) are Eqs. 12 and 13 read off directly; the ratio bound in Eq. 16 is monotone algebra. For (iii): in Eq. 17, $f(u) = Su^2 + N_T(1-u)^2 + N(1-u^\rho)^2$ has $f'(1) = 2S > 0$ and $f'(0^+) = -2N_T < 0$ (for $\rho > 1$; the boundary term vanishes). For $0 < \rho \leq 1$, the right derivative at zero is also negative (or tends to $-\infty$), so a minimizer is interior for every $\rho > 0$. Since $\partial_\rho f = -2N(1-u^\rho)u^\rho \ln u > 0$ there, comparing f_{ρ_1} and f_{ρ_2} at a minimizer of the latter proves strict increase whenever $\rho_1 < \rho_2$, without assuming uniqueness. For $\lambda_N > \lambda_T$, ρ_{nDFA} increases from 1 to ρ_{DFA} as damping increases; when $\lambda_N < \lambda_T$, it decreases from 1 to ρ_{DFA}. The sign claims follow. Eq. 18 is direct minimization at $\rho \in \{1, \infty\}$ ($u^\infty = 0$ for $u < 1$); Eq. 19 compares the two complete minima. □

Interpretation and limitations. In the two-block model, conditioning helps the best point on the trajectory when the task occupies the slower, lower-variance block and can hurt when the task occupies the faster block. This is a statement about the model's rates and early-stopping path, not a guarantee for finite-step nonlinear DFA. In particular, the derivation assumes fixed perfect alignment, $n \geq d$, an invertible empirical covariance, Gaussian label noise, and a population-spectrum approximation for risk. It does not model how feedback alignment is acquired. The full-spectrum crossing and finite-damping expressions below are thus simulation checks of the equations, not independent empirical validation of a network-level law.

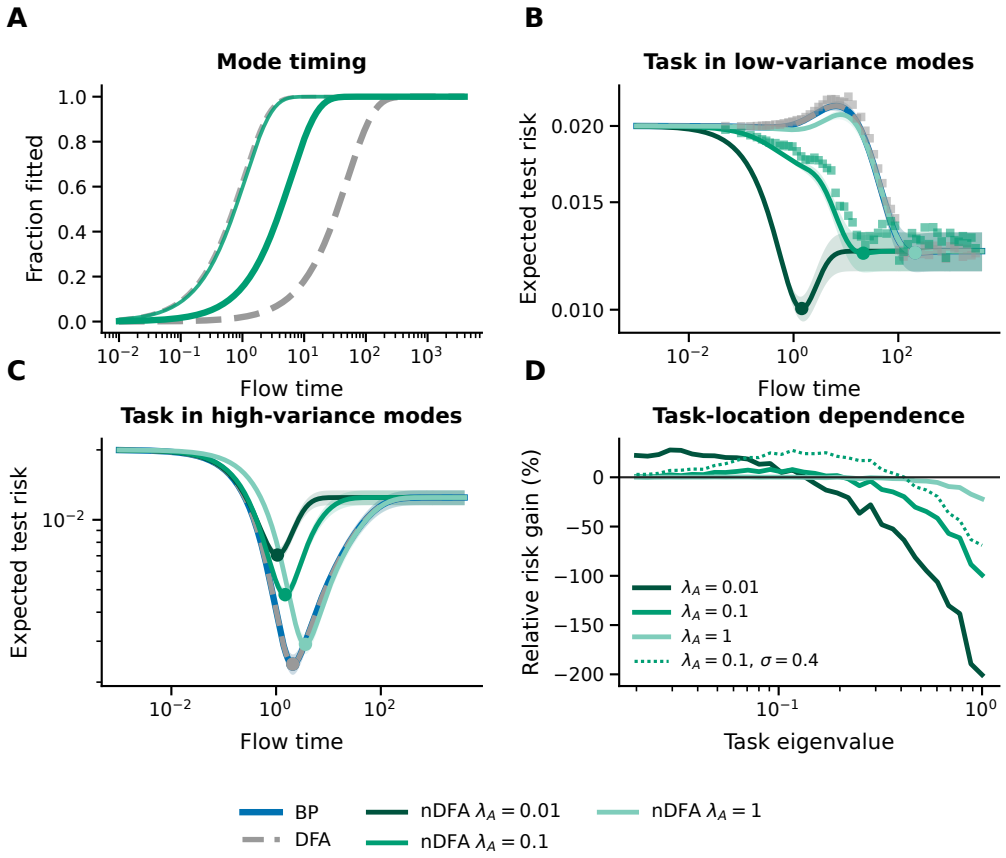

Figure 6: **Mode-timing illustration (Proposition 2). A,** Per-mode fitting curves in the aligned linear model (d=32, κ=50): raw DFA/BP fit the high-variance nuisance mode κ=50× earlier than the low-variance task mode; conditioning compresses the gap to ρ_c=5.5 at λ_A=0.1 while preserving the order. Thin curves mark nuisance and thick curves mark task. **B,** Test-risk trajectories with the task on the two lowest-variance directions (σ=0.2, n=128): raw DFA (grey dashed; coincides with BP, blue) absorbs the nuisance noise before fitting the task and shows no early-stopping gain in this example, while nDFA with $\lambda_A < \lambda_{\text{task}}$ moves the trajectory optimum (dots) lower and, under the chosen clock, earlier; squares: minibatch SGD on the explicit two-layer network. **C,** Task-aligned control: the same re-timing strictly hurts, monotonically in the compression strength—the clean-control reversal. **D,** Best-achievable-risk difference as the task mode sweeps the spectrum: the gain changes sign with task location, and the crossing shifts with σ and λ_A. Bands are ±1 SEM over 20 draws.

Numerical validation. Figure 6 simulates the model exactly (`analysis/validate_mode_timing.py`: d=32, κ=50, log-spaced spectrum, σ=0.2, n=128, S=0.02 on the two lowest-variance directions; both the closed-form gradient-flow solution on the empirical loss and minibatch stochastic gradient descent (SGD) on a two-layer linear network with fixed aligned feedback). The empirical-flow simulation verifies the stated dynamics rather than providing an independent test of their assumptions. Panel A shows compressed mode-timing gaps. In the nuisance-dominant example (panel B), strong conditioning lowers the trajectory minimum (and reaches it earlier under the chosen clock); mild conditioning provides little gain under this signal/noise balance. Minibatch SGD on the explicit aligned network follows the same trend. The task-aligned control reverses the ordering (panel C). Sweeping task location across the spectrum changes the gain's sign, with a crossing that depends on both noise and damping (panel D). The empirical full-spectrum trajectories and the population two-block calculation are distinct models; their quantitative gains need not coincide. The numerical check minimizes all three terms of the two-block risk separately for each rate ratio.

C IMPLEMENTATION, STATISTICS AND STUDY PROVENANCE

The comparisons share a fixed-feedback pathway but differ in normalization, damping and selection. We distinguish these choices before reporting their results; each study states its own departures below. We first distinguish their credit signals. Backpropagation computes

$$\delta_\ell^{\mathrm{BP}} = (W_{\ell+1}^\top \delta_{\ell+1}^{\mathrm{BP}}) \odot \phi'(a_\ell), \qquad \Delta W_\ell^{\mathrm{BP}} = -\eta\, \delta_\ell^{\mathrm{BP}} h_{\ell-1}^\top. \tag{20}$$

Feedback alignment replaces $W_{\ell+1}^\top$ by a fixed random matrix. DFA projects the output error directly as in Eq. 1. DRTP replaces the output error by a random projection of the target. Local auxiliary losses attach supervised heads to hidden layers. VNC and NMNC estimate credit from perturbation-output correlations; NMNC restricts perturbations to a low-dimensional activity manifold.

MLP activity and error factors use full local blocks. Statistics are recomputed from the current minibatch, except for the specified FOOF averages. Fixed Gaussian feedback may be rank-truncated, with its Frobenius norm preserved before applying the feedback scale. The classifier uses the exact output gradient: softmax cross-entropy in the broad benchmarks and one-vs-rest binary log loss in the tanh confirmations. Biases retain their local derivatives. The original MLPs have no BatchNorm unless specified. Diagnostic spectra use centered covariances and should be distinguished from the uncentered moments inverted during training. Damping and solve conventions are specified below.

Algorithm 1: conditioned DFA update for a hidden layer. Given a minibatch, compute DFA hidden errors $\delta_{\ell i} = (B_\ell e_i) \odot \phi'(a_{\ell i})$, where e_i is the derivative of the *unreduced* per-example loss. Form $G_\ell = |\mathcal{B}|^{-1} \sum_i \delta_{\ell i} h_{\ell-1,i}^\top$, $C_A = |\mathcal{B}|^{-1} \sum_i h_{\ell-1,i} h_{\ell-1,i}^\top$, and $C_E = |\mathcal{B}|^{-1} \sum_i \delta_{\ell i} \delta_{\ell i}^\top$. Then choose activity GP_A, error $P_E G$, or K-nDFA $P_E G P_A$, with $P_A = (C_A + \lambda_A I)^{-1}$ and $P_E = (C_E + \lambda_E I)^{-1}$, and multiply by $-\eta$. When the protocol uses norm matching, first rescale the chosen matrix to the Frobenius norm of G_ℓ, after all active factors. The feedback matrix B_ℓ is fixed and random; no transposed forward weights or backpropagated hidden errors are used.

C.1 DIAGNOSTIC DEFINITIONS

To distinguish update direction from magnitude, let g denote a local hidden-weight direction and $g^\star$ its BP counterpart. Inner products and norms are Frobenius products over the weight entries. Then:

$$\cos(g, g^\star) = \frac{\langle g, g^\star \rangle}{\|g\| \|g^\star\|}, \qquad \Pi(g, g^\star) = \frac{\langle g, g^\star \rangle}{\|g^\star\|^2}. \tag{21}$$

The projected ratio measures a first-order descent component in BP units; it depends on magnitude as well as direction and is not a training objective.

Feedback rank measures the number of output-error directions preserved by the broadcast matrix. The activity spectra and nuisance/task energy ratio characterize the geometry on which that feedback acts (Appendix D.1).

C.2 MOMENTS, DAMPING, AND OPTIMIZATION

The scale of a moment determines how strongly its inverse changes an update. We therefore specify the loss normalization and damping conventions before describing the implementation.

For each hidden layer, activities and local errors are obtained from the same minibatch and forward state. The output error is the derivative of the unreduced per-example loss; DFA forms local teaching signals from this error. A local signal computed from a mean loss is multiplied by the actual batch size before forming its second moment. This convention also applies to partial batches. We use uncentered moments unless a control explicitly specifies centering. The A/E/K conditioning operation affects hidden weight matrices only; the separately specified FOOF recipe also conditions the readout. Biases and BN affine parameters retain their local derivatives, while the output classifier receives its exact supervised gradient. Fixed feedback matrices are not optimized.

The original MNIST/Fashion-MNIST studies use absolute damping, with separate values in the units of each moment. The CIFAR-10 studies use

$$\lambda_A = \max\{\rho_A \operatorname{tr}(C_A)/d_{\mathrm{in}}, 10^{-6}\}, \qquad \lambda_E = \max\{\rho_E \operatorname{tr}(C_E)/d_{\mathrm{out}}, 10^{-6}\}, \tag{22}$$

where tr denotes the matrix trace, ρ_A, ρ_E are relative damping coefficients, and $d_{\text{in}}, d_{\text{out}}$ are the layer's input and output widths. Writing G for the raw DFA gradient and U for its conditioned form, we rescale each nonzero U once, after all active factors, by $\widetilde{G} = \|G\|_F U / \|U\|_F$. Matching is to the raw DFA gradient computed at that model's current state, not to a different model's BP gradient. It therefore controls instantaneous hidden-gradient magnitude without equating entire training trajectories.

The original dense implementation symmetrizes moment blocks, solves in float32, imposes a minimum ridge, and can increase the ridge or use least squares after a numerical failure. The larger-network implementation uses exact feature- or minibatch-space solves at the declared ridge, without adaptive ridge escalation or pseudoinverse fallback. Numerical failures remain failed cases rather than silently changing the algorithm. The factor-transfer confirmation reports no such failures. Training uses float32, with higher-precision reductions where specified in the study protocols.

C.3 Exact solves in minibatch space

Conditioning wide layers would be costly if every update required solving a system with one coordinate per unit. The following identities allow the same update to be computed from a smaller system with one coordinate per example when the minibatch is smaller than the layer width.

Let the b rows of $A \in \mathbb{R}^{b \times d_{\text{in}}}$ and $D \in \mathbb{R}^{b \times d_{\text{out}}}$ contain the minibatch activities and local errors. Then $C_A = A^\top A/b$, $C_E = D^\top D/b$, and $G = D^\top A/b$. Define the minibatch Gram matrices $K_A = AA^\top/b$ and $K_E = DD^\top/b$. The push-through identity gives

$$G(C_A + \lambda_A I)^{-1} = b^{-1} D^\top (K_A + \lambda_A I)^{-1} A, \tag{23}$$

$$(C_E + \lambda_E I)^{-1} G = b^{-1} D^\top (K_E + \lambda_E I)^{-1} A, \tag{24}$$

$$(C_E + \lambda_E I)^{-1} G (C_A + \lambda_A I)^{-1} = b^{-1} D^\top (K_E + \lambda_E I)^{-1} (K_A + \lambda_A I)^{-1} A. \tag{25}$$

The two Gram factors need not commute. These identities preserve the current-batch moments and the learning rule; they change only the solve coordinates. The implementation chooses the smaller system separately for each factor. When both use minibatch space, the cost is $O(b^3 + b^2(d_{\text{in}} + d_{\text{out}}) + b\, d_{\text{in}} d_{\text{out}})$, with $O(b^2 + b(d_{\text{in}} + d_{\text{out}}))$ auxiliary storage beyond the parameters and returned gradient. The dense gradient still requires $O(d_{\text{in}} d_{\text{out}})$ storage. These identities apply to current-batch moments, including sample-dependent gates; moving averages require a separate representation or approximation.

C.4 Replication units, selection, and statistical comparisons

Repeated training runs measure how sensitive a result is to training randomness. To interpret that uncertainty, we distinguish independent runs from repeated feedback draws and from changes in the experimental condition.

For the original MNIST and Fashion-MNIST confirmations, the unit of replication is a model/data-order seed after averaging three feedback seeds. Tables report means and standard errors of the mean (SEM). For the additional CIFAR-10 studies, each replicate jointly varies initialization, sampling, augmentation, and feedback as applicable; tables report means and sample standard deviations (SD). Methods are paired within a study and condition. Reusing seed labels across widths or training pools does not make the conditions independent replicates, and earlier development seeds are never pooled with confirmation.

Hyperparameters in these additional studies are selected on validation data. Those earlier studies retain their declared final endpoint. The joint confirmation separately prespecifies validation-CE checkpoint selection for its work-budget comparisons. When the best tested setting lies at the edge of a search grid, we report it as a boundary selection rather than assuming that the optimum was found. The original factor experiments combine separately selected A and E damping in K. The distinct full-pool boundary study uses a joint K grid and is identified separately for this reason.

CIFAR-10 paired intervals are individual two-sided 95% Student t intervals, computed from the seed-level differences. Positive accuracy differences and negative cross-entropy (CE) differences favor the first method. Intervals are conditional on the benchmark, reused split, and development procedure; they do not describe dataset uncertainty or hyperparameter-selection uncertainty. Exact

sign-flip tests additionally assume exchangeability of paired signs. The earlier four-study evaluation prespecifies two accuracy comparisons: A–DFA in the optimizer study and A–FD-DFA, where FD-DFA uses forward decorrelation, in the equal-budget decorrelation study. Holm adjustment uses this fixed family of two. All other comparisons in that earlier evaluation, including factor-transfer contrasts, are descriptive. The later joint confirmation uses the distinct 21-test family in Appendix A.

C.5 Final evaluation and reproducibility scope

Final testing assesses the selected models without using their test scores to revise settings or training duration. The retained prediction records also allow the reported metrics to be checked without retraining.

The first four shared-BN studies evaluate 280 final models: 96 from the scale study, 128 from factor transfer, 24 from optimizer confirmation, and 32 from the forward-decorrelation comparison. The follow-ups and separate boundary study contribute a further 170 and 50 models, respectively (Table 19). The first four studies reproduce their saved validation predictions using training-derived preprocessing and frozen normalization state. No test result selects a model, seed, horizon, or setting. Scale-study A checkpoints are at 5,000 updates; BP/DFA checkpoints are at 8,500. Raw-method checkpoints at 5,000 retain metrics only, so that equal-update comparison is reported on validation.

The official test set was used in earlier project studies. Final model selection was fixed before this evaluation, but the dataset itself was not previously unseen. Prediction-based reproduction checks verify reported accuracy and losses without repeating training or checkpoint inference. Compact records retain exact predicted classes and labels and float64 per-example CE; reconstructing CE directly from logits requires the full archived evidence. Hardware timing and the original order of test access cannot be established from those compact records alone.

C.6 Study protocols and cohorts

C.7 Computational scope

The direct implementation adds a damped moment solve per active factor and layer. For width d and batch size b, forming and solving a dense factor costs $O(d^2b + d^3)$ time and $O(d^2)$ storage. The exact sample-space identities above reduce this cost when the batch is narrower than the layer, using the same low-rank inversion principle as SKFAC (Tang et al., 2021). K-nDFA charges both solves. The measured comparisons in Appendices F–I include these operations and report their timing scopes; equal update counts alone are not equal work.

Tier	Protocol summary
Synthetic stress suite	Eight-class latent-circle task with input dimension 64 and nuisance dimension 24. Regimes are nuisance-dominant $(s_{\mathrm{task}}, s_{\mathrm{nuis}}) = (0.45, 2.0)$, low-sample/noisy $(0.7, 1.5)$, mixed-context $(0.75, 1.2)$ with task–nuisance interaction, and task-aligned $(1.3, 0.25)$. The 128 cells are 4 regimes $\times$ train sizes 512/1024/2048/4096 $\times$ input noise 0.05/0.15 $\times$ train-label noise 0/0.1/0.2/0.4, evaluated on clean test labels with 4096 test examples. MLP hidden widths are 256–128, trained for 14 epochs plus an epoch-0 evaluation, with batch size 128 and learning rate 0.08. The primary comparison fixes full-rank feedback, damping $\lambda = 0.3$, 5 data seeds, and 3 feedback seeds; BP uses the same data seeds. Rank-restricted and test-selected summaries are historical sensitivity analyses only.
Vision MLP noise	Fashion-MNIST and CIFAR-10 MLP sweeps use train sizes 1000/3000/10000 and train-label noise 0/0.1/0.2/0.4, with clean test labels (2000 Fashion-MNIST and 3000 CIFAR-10 test examples in the aggregate). The BP, FA/DFA, DRTP, VNC/NMNC, and nDFA methods are compared; local-feedback methods aggregate over 5 data seeds and 3 feedback seeds, while BP aggregates over data seeds only. The saved sweep fixes full-rank feedback and contains one configuration per method and cell; its common short training recipe is exploratory.
Synthetic controls	BP regularization controls use matched seeds for BP, BP+L2, BP+label smoothing, early stopping, raw DFA, DFA+BP-norm matching, and conditioned DFA.
Clean three-factor MLPs	The tanh studies use three hidden width-300 layers for 1,000 updates (batch 128, learning rate 10^{-3}). A fixed 5,000-example split selects λ_A/λ_E independently: MNIST uses development seeds 42–44, frozen $0.3/10$, and confirmation seeds 50–54; preregistered Fashion-MNIST uses seeds 60–62, frozen $0.03/30$, and confirmation seeds 70–74. The architectural replication uses a 256–128 ReLU MLP with softmax cross-entropy (learning rate 0.03), development seeds 0–2, frozen $3/0.1$, and fresh confirmation seeds 100–107. All cross model seeds with three feedback seeds and norm-match conditioned hidden gradients to DFA. Their later runners evaluate test at the last update; the earlier MNIST test exposure is disclosed in Appendix E.

Table 18: **Protocol summary.** Compact details for the main experimental tiers; code availability is stated in the Reproducibility Statement.

Study / artifact ID	Development and selection	Final cohort	GPU
Scale / S1	Earlier validation choices transferred	8 seeds, 4 cells, 3 rules; 96 models	H100
Error transfer / S2	2 seeds, 7 E dampings; mean CE	8 seeds, 4 cells, 4 rules; 128 models	H100
Optimizer / S3	2 seeds, 3 rates per rule; mean CE	8 seeds, 3 rules; 24 models	H100
Decorrelation / S4	2 seeds, 9 settings per rule; mean CE	8 seeds, 4 rules; 32 models	H100
Longer work / S5	S4 settings transferred	5 seeds, 2 budgets, 4 rules; 40 models	H100
Width retuning / S6	2 seeds, 4 dampings, 2 operators and widths; mean CE	5 seeds, 2 widths, 5 rules; 50 models	H100
BN intervention / S7	Fixed common recipe	5 seeds, 4 conditions, 4 rules; 80 models	H100
Separate protocol / S8	24 trials per rule; final accuracy, CE tie-break	5 seeds, 10 rules; 50 models	H200

Table 19: CIFAR-10 study ledger. Each row is a separate cohort. S1–S4 account for the original 280 final models; S5–S7 add 170, and S8 adds 50. The artifact IDs identify distinct study cohorts; exact configurations and source hashes are retained for the planned code release. Selection uses validation data; confirmation reports the declared final test endpoint. CE denotes cross-entropy. H100 measurements use the 80GB HBM3 model.

D SYNTHETIC EVIDENCE AND MECHANISM CONTROLS

The synthetic suite tests whether conditioning helps when high-variance directions contain nuisance. Table 20 gives the full-rank endpoints; the paired analysis below treats each global seed, rather than each designed cell, as a replicate.

Regime	BP	DFA	nDFA	Best NC	Δ nDFA	Wins
Nuisance-dominant	27.9	13.4	53.3	26.8	+39.9	32/32
Low-sample/noisy	51.4	34.8	65.7	50.0	+30.9	32/32
Mixed-context	38.0	25.3	47.3	36.2	+22.0	32/32
Task-aligned	92.0	74.5	90.4	91.7	+15.9	32/32

Table 20: Primary fixed-full-rank synthetic summary at the shared learning rate. BP comes from this sweep; the separately test-tuned BP references in Fig. 2 and Table 26 use a different study. "Best NC" is the best noise-correlation baseline in the archived sweep. Cells are designed conditions rather than independent population samples.

Regime	BP loss	DFA loss	A loss	DFA loss > 10 (%)
Nuisance-dominant	0.77	86,170.82	0.41	100.0
Low-sample/noisy	0.68	1,350.76	0.38	97.1
Mixed-context	0.93	1,195.00	0.52	100.0
Task-aligned	0.79	1.66	0.60	25.0

Table 21: **Predictive instability in the shared-rate synthetic sweep.** Median final training loss and the fraction of raw-DFA runs with loss above 10. BP uses 160 runs per regime; local rules use 480 crossed data/feedback runs. These distribution summaries do not treat crossed runs as independent replications. No activity-nDFA run exceeds the threshold. Large accuracy differences in this sweep therefore combine stabilization with directional effects; norm-matched controls address the latter separately.

The suite crosses designed conditions with five global data seeds and repeated feedback draws. We average the full designed grid within each global seed before estimating uncertainty. Crossed feedback draws are not additional independent replicates. Table 22 reports this conditional seed-level comparison.

Regime	Comparison	Mean Δ (pp)	95% seed t interval	Wilcoxon p (seeds)
Nuisance-dominant	nDFA − DFA	$+39.86$	$[+37.42, +42.29]$	0.0625
Low-sample/noisy	nDFA − DFA	$+30.88$	$[+27.85, +33.90]$	0.0625
Mixed-context	nDFA − DFA	$+22.02$	$[+21.45, +22.59]$	0.0625
Task-aligned	nDFA − DFA	$+15.85$	$[+13.21, +18.50]$	0.0625

Table 22: Seed-level comparison over the fixed synthetic stress grid. Differences are averaged across all 32 designed cells and retained feedback draws within each of five global data seeds. Individual Student t intervals use four degrees of freedom and assume approximately normal, independent seed-level differences. Cells and crossed feedback draws are not independent replicates; five seeds limit the assessment of coverage. The exact two-sided signed-rank test cannot fall below 0.0625 with five nonzero pairs; all five observed differences have the same sign in every regime.

All five seed-level differences favor activity nDFA in every regime. The small number of seeds limits inference: a two-sided signed-rank test cannot reach the conventional threshold with five nonzero pairs. The separately tuned BP stars in Fig. 2 are descriptive references, not part of this paired comparison.

D.1 PREDICTED ANISOTROPY GAIN

The linearized prediction depends on whether high-variance directions carry nuisance or task signal. Hidden spectra remain anisotropic even in the task-aligned control, so the trained condition number

alone does not identify the favorable regime. The designed nuisance/task energy ratio instead describes how the synthetic generator distributes variance before training. Across 128 cells it correlates with the activity-nDFA gain over DFA (Spearman $\rho = 0.62$; Fig. 4B). This calculation and Table 23 use the same exploratory, test-selected feedback-rank summaries; the primary fixed-full-rank comparison is reported separately above. Because regimes determine much of the ratio, the association is a descriptive mechanism check, not an independent-task significance test or a validated predictor for a new task.

D.2 A PROSPECTIVE ESTIMATE OF THE MECHANISM VARIABLE DOES NOT TRANSFER

The `nuisance_energy_ratio` above is computed from the synthetic generator's ground truth, so it cannot be evaluated before training on a new problem. We tested a purely *prospective* version, estimable from the training data and a randomly initialized network alone: for each presynaptic layer that nDFA preconditions, we estimate the activity covariance C from an untrained forward pass and let U have orthonormal columns spanning the centered class-conditional mean activations. We define $\widehat{r} = (\operatorname{tr} C - \operatorname{tr} U^\top C U)/\operatorname{tr} U^\top C U$, combining layers and five initialization seeds by geometric mean. An empty task subspace has zero projected energy; the implementation floors the denominator at 10^{-12}. The task-blind damped condition number $\kappa(C)$ serves as a baseline (`analysis/compute_prospective_diagnostic.py`, artifacts in `results/infodfa_prospective_diagnostic_v1/`).

The pre-training estimate does not provide a reliable decision rule (Table 23). Its synthetic association is weaker than that of the designed ratio and no better than the task-blind condition number, and the association reverses on the vision transfer. These results limit the proposed diagnostic: the tested initialization statistics do not capture enough of the subsequent learning dynamics to predict when conditioning will help.

Table 23: Prospective nuisance-energy diagnostic vs. realized nDFA−DFA gain. Correlations are Spearman ρ. The pre-training estimate $\widehat{r}$ does not exceed the task-blind $\kappa(C)$ baseline on the synthetic grid, and reverses sign on the held-out vision transfer.

Scope	$\widehat{r}$ (prospective)	$\kappa(C)$ (task-blind)	designed ratio (post hoc)
Synthetic, all 128 cells	0.54	0.59	0.62
Vision transfer, all 24 cells	−0.61	–	–

D.3 CONDITIONING AND FEEDBACK-ALIGNMENT DYNAMICS

The post-alignment model assumes useful credit and does not explain how it is acquired. Historical synthetic alignment trajectories used a shared learning rate and did not separate stabilization from acquisition of useful credit. They therefore do not establish an alignment-rescue mechanism. Gradient alignment alone is also insufficient for good learning performance (Chu and Bacho, 2023); initialization can change the order in which components are learned in linear FA (Girotti et al., 2021). The constant-rate and shared-checkpoint tests in Appendix A provide the controlled timing evidence.

D.4 REDUCED VARIABILITY ACROSS FEEDBACK AND ORDER SEEDS

To assess sensitivity to random feedback, we calculate the within-data-seed SD of final accuracy across three compound feedback/order seeds. The analysis covers 156 synthetic, vision-MLP and MLP-Mixer cells, with 35 chance-floor DFA cells excluded from ratios because their variance is degenerate. Among the remaining cells, the median DFA-to-nDFA SD ratio is 6.1. The reduction appears across regimes and persists descriptively after accounting for mean accuracy (Fig. 7).

These are descriptive summaries of a reused grid, not independent-cell inference. The feedback seed also controls minibatch order, so the result does not isolate feedback-matrix variability. The four MLP-Mixer cells, two at floor, provide little separate evidence. Distinguishing feedback and order effects requires an experiment that varies them independently.

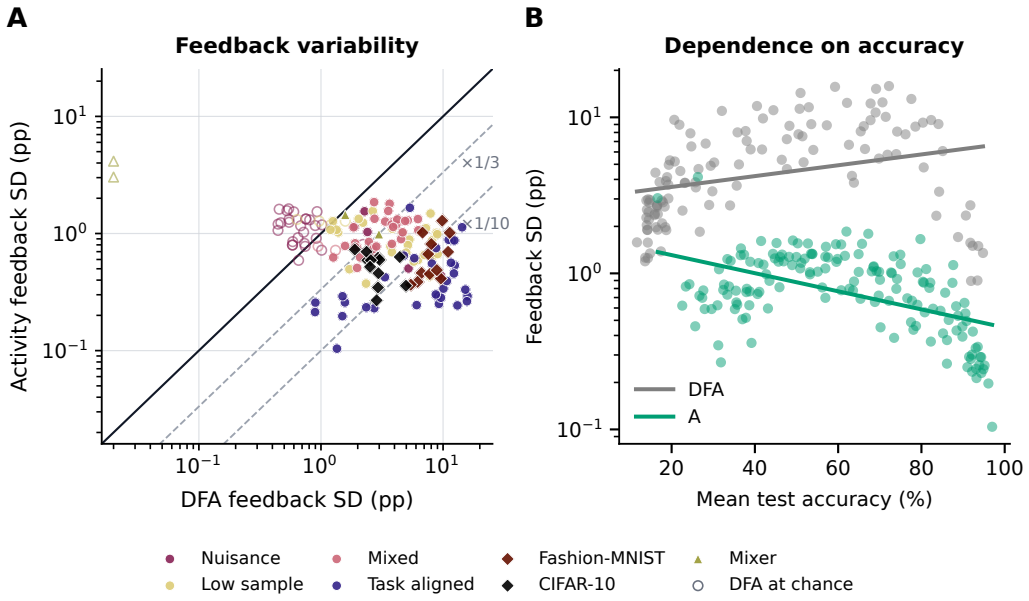

Figure 7: **Variability across compound feedback/order seeds. A,** Per-cell feedback-seed standard deviation of final accuracy, raw DFA versus nDFA (log-log; diagonal = no change), colored by regime; vision and MLP-Mixer cells are filled diamonds/triangles, and cells where raw DFA sits at the chance floor are open circles (excluded from the ratio statistics). Guide lines mark $\times\frac{1}{3}$ and $\times\frac{1}{10}$ reductions. **B,** Ceiling control: feedback-seed standard deviation versus mean final accuracy for above-chance cells. Fitted method-specific trends show that the activity-nDFA reduction persists after accounting descriptively for mean accuracy. Only the consistently normalized raw and activity rules are included.

D.5 Regularization, scalar scaling and moment controls

These controls ask whether stronger BP regularization, a scalar update rescaling, or coordinatewise normalization explains the activity gain. They retain their own matched cohorts and selection procedures; their absolute accuracies should not be pooled across protocols.

Normalization has long been important in learning with asymmetric feedback (Liao et al., 2016). We therefore compare activity conditioning and raw DFA with BN present in both rules, rather than attributing the entire unnormalized gain to the inverse moment.

Task	Method	Test acc	Δ BP	Δ DFA
Synthetic nuisance	BP	58.28 ± 3.16	+0.00	+39.50
	BP + L2 (λ=1e-3)	59.29 ± 3.09	+1.01	+40.51
	BP + label smoothing (0.1)	54.90 ± 3.68	-3.38	+36.12
	BP + early stop	59.22 ± 3.53	+0.94	+40.44
	DFA	18.78 ± 3.42	-39.50	+0.00
	DFA + norm match	12.22 ± 0.41	-46.06	-6.56
	nDFA	72.45 ± 0.62	+14.17	+53.67
Fashion-MNIST noisy	BP	79.19 ± 0.53	+0.00	+17.56
	BP + L2 (λ=1e-3)	79.52 ± 0.35	+0.33	+17.89
	BP + label smoothing (0.1)	79.85 ± 0.28	+0.66	+18.22
	BP + early stop	79.70 ± 0.53	+0.51	+18.07
	DFA	61.63 ± 1.56	-17.56	+0.00
	DFA + norm match	77.00 ± 1.35	-2.19	+15.37
	nDFA	76.49 ± 0.58	-2.70	+14.86

Table 24: **Regularization and norm-matching controls.** Three BP regularization variants test whether a stronger BP recipe closes the gap to conditioned DFA. `DFA + norm match` rescales each layer's DFA gradient to match the BP gradient norm on a held-out evaluation batch, isolating the contribution of activity-side second-moment preconditioning (nDFA) from per-layer norm matching. These accuracies come from a single nuisance cell and a single Fashion-MNIST cell run at five matched seeds with a tuned recipe, and are therefore not directly comparable to the sweep-averaged per-regime accuracies of Table 20, which aggregate 32 cells per regime under the shared hyperparameter protocol of Table 18.

Cell	Method	Test acc.	Δ vs. DFA+norm
Nuisance	DFA	15.3 ± 0.6	+0.2
	DFA + norm	15.1 ± 0.3	+0.0
	activity nDFA + norm	30.4 ± 0.8	+15.3
Low sample/noisy	DFA	23.0 ± 0.8	+1.4
	DFA + norm	21.6 ± 0.6	+0.0
	activity nDFA + norm	31.9 ± 0.4	+10.3
Mixed	DFA	34.6 ± 1.0	+3.3
	DFA + norm	31.2 ± 1.0	+0.0
	activity nDFA + norm	35.4 ± 1.0	+4.1
Clean aligned	DFA	97.2 ± 0.1	+0.5
	DFA + norm	96.7 ± 0.1	+0.0
	activity nDFA + norm	96.4 ± 0.2	-0.3

Table 25: **Norm-matched factor-control endpoints.** Values are final test accuracy in percent for the focused 100-epoch synthetic cells used in Fig. 4. The norm-matched conditioned rows rescale each hidden local gradient to the matched BP layer norm after applying the conditioner.

Regime	BP†	BP+precond.†	Δ	DFA	DFA+$C^{-1/2}$ ‡	nDFA
Nuisance-dominant	27.9	46.2	+18.3	13.5	47.5	53.3
Low-sample/noisy	53.3	60.4	+7.1	34.7	62.0	65.7
Mixed-context	43.8	44.6	+0.72	25.4	46.7	47.5
Task-aligned	92.0	91.9	−0.1	75.0	91.0	90.4

† best of five learning rates per regime, selected by mean test accuracy. ‡ DFA, DFA+decorrelation, and the displayed nDFA column all come from the same matched replication of the 128-cell suite; decorrelation applies $(C + \lambda I)^{-1/2}$.

Table 26: **BP-preconditioning control (best learning rate per method).** Applying the same input-side inverse-second-moment preconditioner used by nDFA to the *exact* BP gradient. BP and BP+precond. are each tuned to their best of five learning rates ($\eta \in \{0.02, 0.04, 0.08, 0.16, 0.32\}$) per regime using test accuracy; the local rules use the fixed sweep rate and receive no analogous rate search. The gain is regime-specific: large on nuisance-dominant, small on mixed-context (where tuning BP's own learning rate recovers most of the apparent gain), and ≈ 0 on the clean task-aligned control. The DFA, decorrelation, and nDFA columns are the matched replication backing the main-text decorrelation deltas, not the separate split-table aggregates. Values are final test accuracy in percent.

Cell	DFA	Hidden Adam	Diagonal $C_A^{-1/2}$	Activity nDFA
Nuisance	17.28 ± 1.19	21.03 ± 1.04	21.26 ± 1.22	45.53 ± 0.93
Low-sample	33.35 ± 0.94	32.33 ± 0.92	33.58 ± 0.95	40.66 ± 0.79
Mixed	37.80 ± 1.92	36.92 ± 1.29	38.78 ± 1.66	38.93 ± 1.36
Task-aligned	97.09 ± 0.17	97.05 ± 0.19	97.10 ± 0.19	96.98 ± 0.14

Table 27: **Exploratory optimizer and diagonal controls.** Test accuracy (%), mean ± SEM over five model seeds after averaging two feedback draws per seed. These are single cells with configurations selected on test accuracy, not a validation-selected optimizer benchmark. The diagonal rule uses inverse square roots while activity nDFA uses an inverse; this comparison does not isolate off-diagonal structure at a common power. Adam adapts hidden-weight updates. The original settings and explicit cohort summaries are retained for the code release.

Regime	DFA+BN	A+BN	Δ (pp)	Individual 95% CI
Nuisance-dominant	47.37	50.72	+3.34	$[+2.58, +4.11]$
Low-sample/noisy	66.65	67.64	+0.99	$[+0.26, +1.71]$
Mixed-context	50.72	48.24	−2.48	$[-3.10, -1.86]$
Task-aligned	87.46	86.48	−0.97	$[-1.77, -0.18]$

Table 28: **Conditioning within a shared BatchNorm recipe.** Regime means of test accuracy (%). Both rules use BN before the hidden ReLU under the synthetic sweep's shared learning rate and 14-epoch horizon. Contrasts average all 32 designed cells and three feedback draws within each of five global seeds; intervals use Student t with four degrees of freedom. Conditioning adds a benefit in the nuisance and low-sample regimes but reduces accuracy in the other two. This comparison does not establish the ordering after independent learning-rate tuning or convergence.

Dataset	BP	DFA	nDFA	NMNC	Δ nDFA	Wins
Fashion-MNIST	72.5	50.2	71.5	73.2	+21.3	12/12
CIFAR-10	32.2	17.2	32.1	30.3	+14.9	12/12

Table 29: Exploratory noisy-label vision MLP sweep. Feedback rank is fixed at full rank; NMNC is the named comparator, not a per-cell selection across noise-correlation methods. Training uses the common short recipe.

D.6 Transfer to a spurious-correlation task

Our ColoredMNIST-style task tests transfer under a controlled change in nuisance structure (Arjovsky et al., 2019). The training color–label correlation is reversed at test, and a grayscale probe removes color. Conditioning reduces seed variability and brings mean accuracy close to BP (Table 30). The paired nDFA–DFA accuracy gains do not survive Holm correction at eight seeds. This supports a descriptive stability result, without establishing superiority to DFA or BP.

Method	Test (color-reversed)	Grayscale	Δ DFA (test)	Δ BP (test)	Δ DFA (gray)
BP	82.81 ± 0.72	91.24 ± 0.22	+8.57	–	+10.88
BP + L2	82.73 ± 0.80	91.23 ± 0.22	+8.49	-0.09	+10.87
DFA	74.24 ± 4.64	80.36 ± 4.35	–	-8.57	–
DFA + norm match	72.29 ± 0.27	85.46 ± 0.27	-1.94	-10.52	+5.10
nDFA	82.50 ± 0.32	91.87 ± 0.18	+8.26	-0.31	+11.51

Table 30: ColoredMNIST-style spurious-correlation task, 8 seeds. One training environment has 85% label/color correlation, test has 15%, and labels remain clean. This differs from the two training environments and label noise of Arjovsky et al. (2019). The grayscale probe removes color. Values are mean ± SEM over eight paired training seeds. Differences are descriptive: the nDFA–DFA improvements do not survive Holm correction over the tested comparison family.

E Independent activity and error confirmations

To test the two factors separately, we use three width-300 tanh hidden layers, sigmoid outputs, and one-vs-rest binary log loss. Training uses 1,000 updates, batch size 128, and learning rate 10^{-3}. Figure 8A–B shows independent damping selection, and panel C shows the frozen confirmation of Eq. 3. Development uses model/data-order seeds 42–44 and one feedback seed. Activity and error damping are selected independently from $\{0.03, 0.1, 0.3, 1, 3, 10\}$ on a fixed 5,000-example split of the MNIST training set. The selected values are $\lambda_A = 0.3$ and $\lambda_E = 10$; K-nDFA simply combines them. The error choice lies at the upper grid boundary and validation accuracy was still increasing there, so the MNIST error damping is not fully localized. Confirmation freezes these choices and uses model/data-order seeds 50–54 crossed with three feedback seeds. We average feedback seeds within each model seed before computing the five paired differences. This later runner evaluates test only at its final step. However, an earlier runner had already used the MNIST test set for repeated evaluation during training and evaluated activity and K-nDFA on seeds 50–54 before validation development. Thus these seeds are not a test-blind confirmation, and subsequent validation selection does not undo the earlier exposure. The separately preregistered Fashion-MNIST and ReLU studies provide the confirmatory replications; the later MNIST seed extension is reported separately as post-hoc evidence.

The result supports the algebraic symmetry while also showing an empirical asymmetry. Activity conditioning has the larger standalone effect and tolerates moderate damping. Error conditioning has a smaller but independent benefit and requires substantially heavier damping under this recipe. Fixed feedback and tanh saturation shape the local errors, but these experiments do not isolate error anisotropy as the cause of the damping dependence. K-nDFA is useful when both corrections are active; it should not be interpreted as automatically superior in regimes where C_E is noisy or nearly singular.

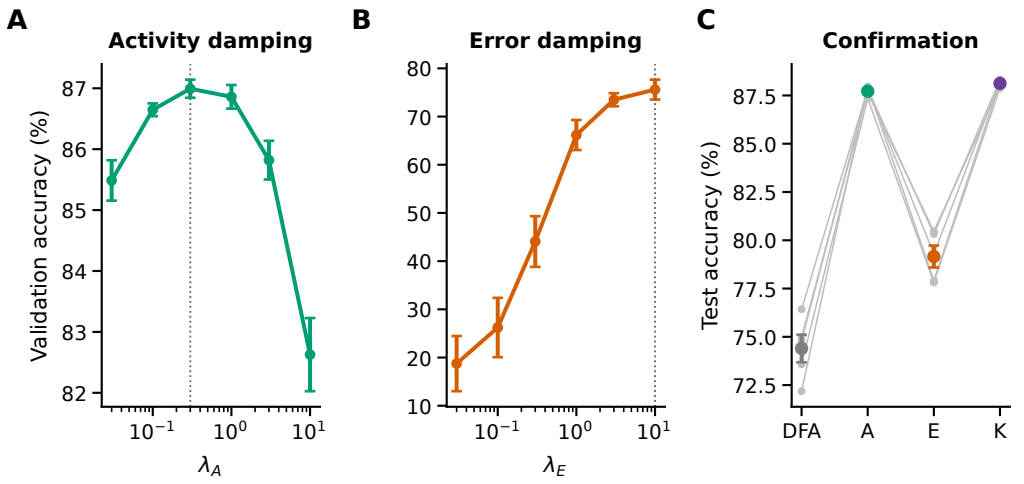

Figure 8: **Activity, error, and two-sided conditioning. A–B,** Development validation accuracy selects activity damping λ_A=0.3 and the much heavier error damping λ_E=10 independently; whiskers are SEM over three development seeds. **C,** Frozen test confirmation. Gray lines connect the four methods within each model/data-order seed after averaging three feedback seeds; colored points show seed means ± SEM. Error nDFA improves raw DFA without injected input noise, while K-nDFA improves both single-sided rules. All conditioned hidden gradients are norm-matched to raw DFA.

Method	Activity factor	Error factor	Test accuracy (%)	Test loss
DFA	–	–	74.40 ± 0.71	1.844 ± 0.013
activity nDFA	✓	–	87.72 ± 0.11	1.272 ± 0.007
error nDFA	–	✓	79.16 ± 0.57	1.620 ± 0.013
K-nDFA	✓	✓	$\mathbf{88.12 \pm 0.09}$	$\mathbf{1.206 \pm 0.004}$

Table 31: Frozen three-factor confirmation, mean ± SEM over five model/data-order seeds after averaging three feedback seeds within each. The paired accuracy gains are +13.32 pp for activity nDFA over DFA, +4.76 pp for error nDFA over DFA, and +0.40 pp for K-nDFA over activity nDFA. Every contrast agrees in sign across all five seed-level pairs; the two-sided Wilcoxon test is floor-limited at $p = 0.0625$.

E.1 Preregistered Fashion-MNIST replication and source-swap diagnosis

Before reading any development endpoint, we registered the clean Fashion-MNIST protocol and pass/fail criteria in a timestamped protocol. The architecture, optimizer, loss, step budget, norm matching, and train/validation/test discipline match the MNIST study, but the split and model/data-order seeds are fresh. Development seeds 60–62 select λ_A and λ_E independently (Table 32) from $\{0.03, 0.1, 0.3, 1, 3, 10, 30, 100\}$; confirmation freezes the selected $0.03/30$ pair on seeds 70–74 crossed with three feedback seeds. The activity optimum lies at the lower grid boundary, so its absolute damping is not fully localized; the error optimum is interior.

The frozen confirmation supports the error-only and incremental two-sided gains (Fig. 3C). The BP-error comparator does not isolate an error-source effect: its moment is much smaller than the local DFA-error moment, so the shared damping makes it nearly scalar. Norm matching then makes its update almost identical to activity nDFA. An audited trajectory confirms this scale mismatch. The source-swap criterion therefore provides no evidence for source equivalence or the absence of a locality penalty.

A post-hoc validation sweep of BP-source damping over $\{0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1, 3, 10, 30\}$ selects 3, with loss breaking an accuracy tie. Five fresh model/data-order seeds (80–84), each averaged over three feedback draws, again favor local K-nDFA over activity nDFA. The retuned BP-source factor adds little benefit and remains almost parallel to the activity-only update on the audited trajectory. This comparison is exploratory

Method	Activity factor	Error source	Test accuracy (%)	Test loss
DFA	–	–	61.98 ± 0.17	1.938 ± 0.012
activity nDFA	✓	–	80.29 ± 0.10	1.575 ± 0.007
error nDFA	–	DFA	63.74 ± 0.41	1.842 ± 0.010
K-nDFA	✓	DFA	$\mathbf{80.79 \pm 0.09}$	$\mathbf{1.500 \pm 0.007}$
K-nDFA (BP source)	✓	BP	80.29 ± 0.10	1.576 ± 0.007

Table 32: Frozen Fashion-MNIST confirmation, mean ± SEM over five model/data-order seeds after averaging feedback seeds. The preregistered contrasts are $+1.77$ pp for error nDFA over DFA, $+0.50$ pp for K-nDFA over activity nDFA, and $+0.499$ pp for local K-nDFA over the BP-source comparator; each has the same sign in all five pairs.

and specific to the tested source scales; it does not establish source equivalence. The ReLU/softmax study below tests a separate question: whether the two factor gains persist under a different activation and loss.

E.2 Fresh-seed ReLU/softmax architectural replication

The next confirmation changes the tanh DFA-stall architecture and one-vs-rest binary log loss simultaneously. A 256–128 ReLU MLP uses multiclass softmax cross-entropy, learning rate 0.03, batch size 128, and 1,000 updates. Development seeds 0–2 select λ_A from $\{0.03, 0.1, 0.3, 1, 3, 10, 30\}$ and λ_E from $\{0.003, 0.01, 0.03, 0.1, 0.3, 1, 3, 10\}$ on a fixed 5,000-example split. The selected $3/0.1$ pair is interior to both grids (Table 33); panel C reports the test confirmation. We recorded the protocol and pass/fail criteria before test evaluation. Confirmation uses fresh model/data-order seeds 100–107 crossed with feedback seeds 0–2, averages feedback seeds within model seed, and evaluates test only after the last update. As in the tanh studies, conditioned hidden weight gradients are norm-matched to raw DFA.

Method	Activity factor	Error factor	Test accuracy (%)	Test loss
DFA	–	–	87.08 ± 0.79	187.339 ± 95.506
activity nDFA	✓	–	95.81 ± 0.03	0.1385 ± 0.0005
error nDFA	–	✓	94.61 ± 0.16	1.242 ± 0.187
K-nDFA	✓	✓	$\mathbf{96.71 \pm 0.04}$	$\mathbf{0.1103 \pm 0.0009}$

Table 33: Frozen ReLU/softmax confirmation, mean ± SEM over eight model/data-order seeds after averaging three feedback seeds. Error nDFA improves DFA by $+7.53 \pm 0.86$ pp; K-nDFA improves activity nDFA by $+0.90 \pm 0.04$ pp and lowers loss by 0.0282 ± 0.0011. Both accuracy contrasts have the predicted sign in 8/8 pairs (two-sided Wilcoxon $p = 0.0078125$).

Raw DFA's unusually large cross-entropy reflects rare confident errors from unstable logits rather than a mismatch in the reported accuracy; all methods use the same softmax objective and evaluation. Error-only conditioning sharply reduces that instability, while the two-sided rule gives the best accuracy and loss. This confirmation supports architecture/loss transfer on MNIST. It does not erase a negative scope result: in exploratory ReLU/softmax Fashion-MNIST development, independently selected factors did not preserve an incremental K-nDFA gain. We therefore do not claim ReLU dataset generality.

E.3 Post-hoc seed extensions

The frozen protocols were later repeated with additional model seeds. Both contrasts retain their sign in the added cohorts (Table 34). These extensions improve the precision of the descriptive comparison but do not address the short training horizon, establish tuned-baseline superiority, or remove earlier test exposure. The original cohorts remain the basis of the original tables and figures.

Setting	Cohort	Seeds	n	E−DFA	K−A	Nominal p
MNIST/tanh	Extension	55–59	5	+4.21	+0.38	0.06250
MNIST/tanh	Pooled	50–59	10	+4.48	+0.39	0.00195
Fashion/tanh	Extension	75–79	5	+1.54	+0.46	0.06250
Fashion/tanh	Pooled	70–79	10	+1.65	+0.48	0.00195
MNIST/ReLU	Extension	108–112	5	+9.05	+0.91	0.06250
MNIST/ReLU	Pooled	100–112	13	+8.11	+0.91	0.00024

Table 34: **Post-hoc extensions of the frozen factor protocols.** Mean paired accuracy differences in percentage points. Each model/data-order seed averages three feedback draws. Every seed favors E over DFA and K over A; the last column is the identical exact two-sided Wilcoxon value for the two contrasts, without multiplicity adjustment. Pooled results include the original cohort and the post-hoc extension; they are descriptive and do not turn the original MNIST cohort into a test-blind confirmation. Explicit original, extension and pooled exports prevent the larger cohorts from silently replacing the original tables and figures.

F CIFAR-10: ACTIVITY GEOMETRY AND TRANSFER ACROSS SCALE

We next ask whether the activity-side gain survives forward normalization and extends to more training data and wider networks. The studies share the forward and data protocol below, then vary training-pool size, network width, and computational budget. Here A, E, and K abbreviate activity-only, error-only, and two-sided nDFA; CE denotes cross-entropy loss.

F.1 SHARED FORWARD AND DATA PROTOCOL

Holding normalization and data processing fixed helps separate the effect of conditioning from differences in the forward network.

The shared-BN studies reserve 2,000 of the 50,000 training images for validation (split seed 80423). Training uses either the remaining 48,000 identities or a nested 10,000-image pool (data seed 9173). Channel means and population standard deviations are fitted separately on each training pool. Each update samples 64 identities with replacement and creates two independently cropped and horizontally flipped views of each. Inputs are flattened for a ReLU MLP with hidden widths 1024–512 or 2048–1024. Including BN affine parameters, these models have 3,679,754 and 8,408,074 supervised parameters. BN precedes each hidden ReLU; running statistics are frozen at validation and test. Feedback scale is 0.1.

The initial activity study selects plain-SGD learning rates of 0.3 for BP and 0.1 for DFA/A, and relative activity damping $\rho_A = 30$. These settings are transferred unchanged in the scale and factor studies. The initial search has unequal method-specific trial counts and selected boundary settings; it is not used to claim equally tuned performance. The subsequent matched-work studies address search and optimizer fairness with separate fresh-seed comparisons.

F.2 LARGER TRAINING POOLS, WIDER MODELS, AND LONGER BASELINES

To test transfer without further tuning, we carry the selected settings across training-pool sizes and widths. Longer unconditioned runs assess how much of the equal-update advantage can be recovered with extra training.

The scale confirmation uses seeds 33000–33007 in each of the four training-pool/width cells. A trains for 5,000 updates, while BP and DFA continue to 8,500; their 5,000-update metrics come from those same trajectories. The earlier development runs are not pooled into confirmation. The principal equal-update contrast is A–DFA at 5,000 updates.

On validation, every pair favors A in both accuracy and CE at equal updates. The advantage persists across the tested pool sizes and widths, while longer raw-DFA training reduces it. In both larger-pool cells the longer-DFA accuracy interval spans zero, and longer BP outperforms A. Increasing width enlarges the A–DFA gap by degrading DFA more than A; it does not improve A's absolute accuracy.

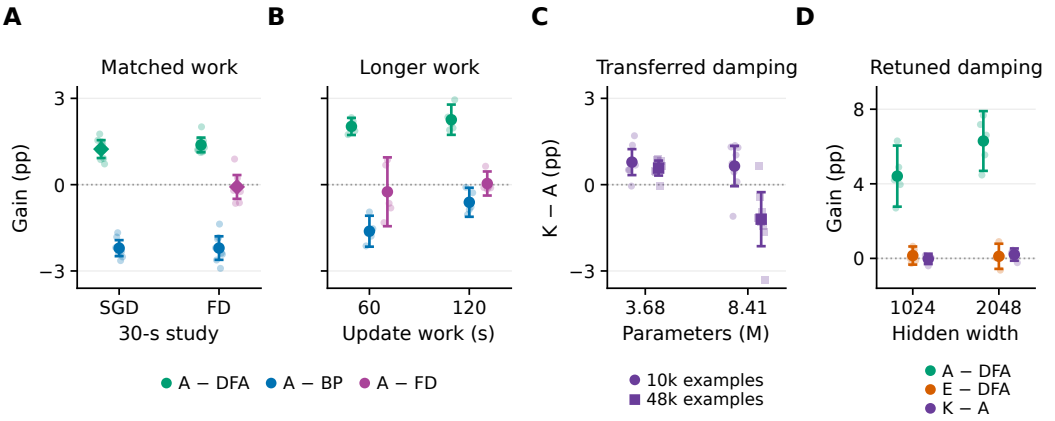

Figure 9: **Earlier CIFAR-10 work and width studies. A,** A versus DFA, BP, and forward-decorrelated DFA (FD), in separate optimizer (SGD) and matched-FD studies at 30 seconds of measured update work. **B,** Longer matched-work budgets retain the activity gain over DFA; BP remains stronger and FD is competitive. **C,** K versus A at 5,000 updates with transferred damping, across parameter counts and training-pool sizes (circles: 10k; squares: 48k). **D,** Width-specific validation selection retains the activity gain but provides little additional error-factor benefit. Faint points show paired seed differences; large markers and whiskers show means and individual 95% t intervals (eight seeds in A and C; five in B and D). Diamonds mark the two prespecified primary comparisons in A; the remaining contrasts are descriptive. Cohorts are analyzed separately; different budgets are separate training runs, not checkpoints along a shared trajectory.

Tables 35–37 report measured costs and final test comparisons. The complete validation trajectories are retained in the retained records.

Individual paired intervals are descriptive. The four equal-update accuracy comparisons additionally have exact sign-flip $p = 0.0078125$ and Holm-adjusted $p = 0.03125$ over that family. These tests concern training-seed variation within the fixed conditions, not a pooled set of independent tasks. Measured update times show that the longer raw baselines still use somewhat less work than A, so the extension is a nominal allocation rather than an exact equal-time comparison.

Cell	Raw rule	$t(A_5)/t(raw_5)$	$t(raw_{8.5})/t(A_5)$	95% CI (last ratio)
10k / 3.68M	DFA	1.805 ± 0.067	0.942 ± 0.033	$[0.914, 0.969]$
	BP	1.795 ± 0.044	0.945 ± 0.026	$[0.924, 0.966]$
48k / 3.68M	DFA	1.791 ± 0.010	0.948 ± 0.005	$[0.943, 0.952]$
	BP	1.779 ± 0.011	0.954 ± 0.006	$[0.949, 0.959]$
10k / 8.41M	DFA	1.814 ± 0.012	0.936 ± 0.006	$[0.931, 0.941]$
	BP	1.816 ± 0.012	0.935 ± 0.006	$[0.930, 0.940]$
48k / 8.41M	DFA	1.810 ± 0.012	0.938 ± 0.006	$[0.933, 0.943]$
	BP	1.804 ± 0.017	0.941 ± 0.009	$[0.934, 0.949]$

Table 35: Scale-study measured update-time ratios. Ratios are formed within each seed before averaging. The final interval describes the raw 8,500-update time divided by A at 5,000; values below one mean that the longer raw baseline still uses less measured work.

F.3 Full versus diagonal activity moments

A direct control tests whether off-diagonal activity-moment structure helps beyond separate coordinate scales. In an existing CIFAR-10 confirmation, full activity conditioning is compared with $G[\mathrm{diag}(C_A) + \lambda_A I]^{-1}$. Both match the raw DFA hidden update norms, share BN and a 3072–1024–512–10 ReLU MLP, and use the same paired training seeds. Each model trains for 500 updates on 10,000 images. The settings were selected on validation before the eight-seed confirmation: both use

Train	Params (M)	Rule	Test accuracy (%)	Test CE
10k	3.68	BP	53.96 ± 0.50	1.4148 ± 0.0188
10k	3.68	DFA	52.13 ± 0.90	1.4195 ± 0.0206
10k	3.68	A	54.56 ± 0.53	1.3466 ± 0.0171
10k	8.41	BP	54.34 ± 0.61	1.4316 ± 0.0185
10k	8.41	DFA	49.57 ± 1.72	1.5925 ± 0.0726
10k	8.41	A	54.13 ± 0.98	1.4467 ± 0.0494
48k	3.68	BP	58.02 ± 0.48	1.1825 ± 0.0097
48k	3.68	DFA	54.03 ± 0.83	1.2899 ± 0.0202
48k	3.68	A	55.45 ± 1.13	1.2539 ± 0.0296
48k	8.41	BP	58.55 ± 0.31	1.1664 ± 0.0088
48k	8.41	DFA	51.76 ± 1.42	1.4044 ± 0.0608
48k	8.41	A	53.37 ± 2.47	1.3658 ± 0.1146

Table 36: Scale-study final test endpoints. Mean $\pm$ sample SD across eight seeds per cell. A ends at 5,000 updates; BP and DFA end at 8,500.

Cell	Contrast	Accuracy (pp)	95% CI	CE	95% CI
10k/3.68	A−BP	+0.596	$[-0.082, +1.275]$	-0.0682	$[-0.0935, -0.0429]$
10k/3.68	A−DFA	+2.430	$[+1.360, +3.500]$	-0.0729	$[-0.0979, -0.0479]$
10k/8.41	A−BP	-0.211	$[-1.219, +0.796]$	+0.0151	$[-0.0326, +0.0628]$
10k/8.41	A−DFA	+4.563	$[+2.411, +6.714]$	-0.1458	$[-0.2365, -0.0551]$
48k/3.68	A−BP	-2.578	$[-3.539, -1.616]$	+0.0715	$[+0.0455, +0.0975]$
48k/3.68	A−DFA	+1.414	$[+0.388, +2.439]$	-0.0360	$[-0.0638, -0.0081]$
48k/8.41	A−BP	-5.181	$[-7.270, -3.092]$	+0.1994	$[+0.1030, +0.2959]$
48k/8.41	A−DFA	+1.606	$[-0.425, +3.638]$	-0.0386	$[-0.1228, +0.0456]$

Table 37: Activity comparisons at the final test endpoints. Accuracy is in percentage points; CE is unscaled. Positive accuracy and negative CE favor the first method. Individual 95% intervals are descriptive and unadjusted.

learning rate 0.1, with relative damping 30 for the full moment and 0.3 for its diagonal. This compares separately selected operators, rather than imposing identical damping or claiming an exactly balanced search.

Full conditioning reaches $46.51 \pm 1.04\%$ test accuracy and the diagonal control $42.10 \pm 1.34\%$ (sample SD). Every paired seed favors the full moment. The additional full-minus-diagonal contrast is 4.41 percentage points with an individual 95% t interval $[3.71, 5.10]$, conditional on the fixed data split and approximately normal independent seed differences. This contrast was not one of the original five declared tests and is reported as an additional analysis. The result supports useful off-diagonal uncentered moment structure under this recipe. Since $C_A = \mathrm{Cov}(h) + \mu\mu^\top$, with $\mu = \mathbb{E}[h]$, the diagonal control removes both off-diagonal centered covariance and off-diagonal mean products. BN before ReLU does not center the post-ReLU activities. This experiment therefore does not distinguish centered correlations from mean-activity effects, or establish that coordinatewise methods generally cannot reproduce the benefit.

G CIFAR-10: TRANSFER OF THE ERROR FACTOR

The activity-side gain leaves open whether error conditioning adds a benefit at the same scales. We test this by fixing the activity damping, selecting error damping using E alone, and evaluating their combination.

A separate study compares DFA, A, E, and K in every training-pool/width cell, using seeds 35000–35007 and the shared plain-SGD protocol for 5,000 updates. Activity damping remains fixed at $\rho_A = 30$. E-only development on the smaller-model, smaller-pool cell uses seeds 34000 and 34001 to select a single error damping by final mean validation CE. The selected $\rho_E = 30$ is transferred unchanged to E and K in every cell. The grid is $\{0.03, 0.1, 0.3, 1, 3, 10, 30\}$, and K outcomes do

not enter selection. The selected upper boundary leaves the optimum beyond the grid unresolved; development scores are retained for the code release.

The activity-side benefit repeats in every pair. K lowers validation CE relative to A in all pairs of both narrower-model cells, but all final validation-accuracy intervals span zero. Final testing favors K in the narrower cells and favors A in the widest, larger-pool cell. In the latter, the validation trajectory changes from an early K advantage to an endpoint disadvantage. E alone is not a reliable improvement on DFA. These results support conditional complementarity (Fig. 9C), with all factor-test contrasts treated as descriptive. Both factors are charged in the measured cost: K requires more update time than A at the same number of updates.

Cell	$t(A)/t(DFA)$	$t(E)/t(DFA)$	$t(K)/t(A)$	95% CI (last ratio)
10k / 3.68M	1.639 ± 0.033	1.638 ± 0.025	1.130 ± 0.017	$[1.116,\ 1.145]$
48k / 3.68M	1.645 ± 0.025	1.648 ± 0.022	1.140 ± 0.028	$[1.117,\ 1.163]$
10k / 8.41M	1.675 ± 0.024	1.681 ± 0.022	1.132 ± 0.019	$[1.116,\ 1.147]$
48k / 8.41M	1.681 ± 0.010	1.688 ± 0.017	1.130 ± 0.005	$[1.125,\ 1.134]$

Table 38: Factor-transfer measured update-time ratios. Within-seed ratios are averaged over eight seeds; the final column gives the individual interval for K/A. These equal-update costs use the factor runner's timing scope.

Train	Params (M)	Rule	Test accuracy (%)	Test CE
10k	3.68	DFA	50.86 ± 0.62	1.4175 ± 0.0152
10k	3.68	A	55.11 ± 0.56	1.3353 ± 0.0138
10k	3.68	E	51.01 ± 0.43	1.4212 ± 0.0117
10k	3.68	K	55.89 ± 0.33	1.3148 ± 0.0134
10k	8.41	DFA	48.83 ± 1.21	1.5669 ± 0.0589
10k	8.41	A	54.46 ± 0.63	1.4280 ± 0.0321
10k	8.41	E	47.51 ± 1.33	1.6748 ± 0.0945
10k	8.41	K	55.10 ± 0.85	1.4192 ± 0.0532
48k	3.68	DFA	50.96 ± 0.67	1.3770 ± 0.0242
48k	3.68	A	55.70 ± 0.75	1.2514 ± 0.0220
48k	3.68	E	49.84 ± 1.00	1.4169 ± 0.0326
48k	3.68	K	56.28 ± 0.80	1.2385 ± 0.0224
48k	8.41	DFA	48.00 ± 1.02	1.5528 ± 0.0669
48k	8.41	A	53.93 ± 1.27	1.3522 ± 0.0547
48k	8.41	E	45.50 ± 2.24	1.7419 ± 0.1140
48k	8.41	K	52.73 ± 1.92	1.4415 ± 0.0753

Table 39: Complete factor-transfer test endpoints at 5,000 updates. Mean $\pm$ sample SD over eight seeds per cell.

Cell	Contrast	Accuracy (pp)	95% CI	CE	95% CI
10k/3.68	A−DFA	+4.245	$[+3.890, +4.600]$	-0.0822	$[-0.0930, -0.0713]$
10k/3.68	E−DFA	+0.149	$[-0.484, +0.782]$	+0.0037	$[-0.0061, +0.0135]$
10k/3.68	K−A	+0.784	$[+0.332, +1.235]$	-0.0205	$[-0.0266, -0.0144]$
10k/8.41	A−DFA	+5.626	$[+4.898, +6.355]$	-0.1389	$[-0.1746, -0.1032]$
10k/8.41	E−DFA	-1.323	$[-2.101, -0.544]$	+0.1079	$[+0.0528, +0.1630]$
10k/8.41	K−A	+0.646	$[-0.054, +1.346]$	-0.0088	$[-0.0469, +0.0292]$
48k/3.68	A−DFA	+4.736	$[+4.144, +5.328]$	-0.1256	$[-0.1406, -0.1107]$
48k/3.68	E−DFA	-1.120	$[-1.559, -0.681]$	+0.0399	$[+0.0225, +0.0572]$
48k/3.68	K−A	+0.576	$[+0.316, +0.836]$	-0.0129	$[-0.0203, -0.0055]$
48k/8.41	A−DFA	+5.929	$[+5.197, +6.661]$	-0.2006	$[-0.2375, -0.1636]$
48k/8.41	E−DFA	-2.504	$[-4.020, -0.987]$	+0.1891	$[+0.1029, +0.2753]$
48k/8.41	K−A	-1.200	$[-2.139, -0.261]$	+0.0893	$[+0.0566, +0.1219]$

Table 40: Factor comparisons at the final test endpoints. Positive accuracy and negative CE favor the first method. Individual 95% intervals are descriptive and unadjusted; K–A measures the incremental error factor, as in the main figure.

H CIFAR-10: COMPARISONS AT MATCHED MEASURED WORK

Because conditioning increases the cost of each update, equal update counts do not imply equal computational cost. These studies compare methods at the same measured update time and with equal numbers of tuning trials.

H.1 COMMON OPTIMIZER AND WORK BUDGET

A shared optimizer and time-based schedule make the computational comparison interpretable even when methods complete different numbers of updates.

Both matched-work studies use the larger training pool and narrower MLP, shared BN, and momentum SGD with momentum 0.9, no Nesterov acceleration, and weight decay 5×10^{-4} on weight matrices only. Learning rate warms linearly over the first 5% of measured work from 0.1 times the peak, then follows a cosine decay to 0.01 times the peak. The schedule advances by synchronized update time divided by a 30-second budget. Sampling, augmentation, gradient construction, conditioning or decorrelation, and optimization are charged. Validation, loading, export, and full-state diagnostics are excluded; total process time is reported separately. The final update's budget overshoot is included. Both original 30-second studies ran on an NVIDIA H100 80GB HBM3, as did the 60- and 120-second follow-ups. Training tensors are float32 and no mixed-precision autocast is used. The follow-up code explicitly disables TF32 for matmul and cuDNN; the original 30-second receipts do not record these runtime flags. Each budget uses a separate cohort, so the columns do not describe checkpoints from one trajectory or an otherwise identical budget intervention.

H.2 EQUAL LEARNING-RATE SEARCH FOR BP, DFA, AND A

This first comparison asks whether the activity-side gain remains when raw DFA receives its own learning-rate search under the same work budget.

Development seeds 36000–36001 evaluate peak rates $\{0.01, 0.03, 0.1\}$ for every method. Mean final validation CE selects the setting, followed by mean accuracy and then the smaller peak rate. Activity damping is inherited from the earlier study. Selection chooses BP/DFA/A peak rates 0.1/0.03/0.1; BP and A reach the upper boundary. Confirmation uses eight fresh paired seeds, 37000–37007.

A improves DFA in validation and test accuracy and CE across all paired seeds, despite completing fewer updates in the common work budget. The final-test primary comparison remains significant under the declared two-comparison Holm adjustment (Fig. 9A, SGD rows). BP is stronger than A. This establishes an activity-side benefit under the specified finite optimizer search and work budget; boundary selections leave fully tuned convergence unresolved.

Rule	Updates	Measured update seconds	Total child seconds
BP	14429.5 ± 88.4	30.001131 ± 0.000724	43.67 ± 3.39
DFA	14485.1 ± 75.1	30.000759 ± 0.000425	44.69 ± 3.26
A	8032.3 ± 65.5	30.001770 ± 0.001249	43.10 ± 3.45

Table 41: Optimizer-study updates and costs at the work endpoint. Mean ± sample SD over eight seeds. Total process time includes overhead excluded from the update clock.

Train	Params (M)	Rule	Test accuracy (%)	Test CE
48k	3.68	BP	63.52 ± 0.20	1.0328 ± 0.0031
48k	3.68	DFA	60.07 ± 0.26	1.1339 ± 0.0043
48k	3.68	A	61.31 ± 0.29	1.0979 ± 0.0036

Table 42: Optimizer-study test endpoints. Mean ± sample SD across eight paired seeds at the measured-work endpoint.

H.3 Equal-budget comparison with forward decorrelation

Forward decorrelation provides a closely related way to address activity correlations. We compare it with update conditioning while allowing both methods to tune their method-specific parameter.

Development seeds 38000–38001 evaluate nine settings each for BP, DFA, A, and FD-DFA; confirmation uses seeds 39000–39007. BP and DFA search nine peak learning rates from 0.003 to 1. A crosses $\{0.03, 0.1, 0.3\}$ with activity dampings $\{3, 30, 300\}$; FD-DFA crosses the same three peaks with decorrelator adaptation rates $\{10^{-6}, 10^{-5}, 10^{-4}\}$. Mean final validation CE selects one setting per method before confirmation. Candidate-level development scores are retained for the code release. Equal trial counts do not equate coordinate resolution across these different searches.

FD-DFA adapts the dense forward-decorrelation mechanism of Ahmad (2024), pinned to official-code revision `00cf470`. It does not reproduce that work's complete FA/Adam protocol. Dense decorrelators precede every linear map, including the input and classifier maps. During training they center activities, apply the current matrix, then update it once using the reference gain and left-multiplied covariance correction. An algebraically equivalent low-rank multiplication avoids a cubic matrix product. Evaluation freezes matrices and running means. Direct feedback is injected at the original post-ReLU activity before the outgoing decorrelator; gradients then use the local ReLU/BN derivative and decorrelated presynaptic inputs. Independent local derivative checks verify this credit coordinate.

All decorrelator transformations and updates are charged to the work budget. They add stored state during training, although the resulting affine transforms can be folded into the following linear maps for inference. The state count is therefore not a deployment lower bound. Selected settings are peak 0.1 for BP, peak 0.03 for DFA, peak 0.03 with $\rho_A = 3$ for A, and peak 0.03 with adaptation rate 10^{-5} for FD-DFA. A and FD-DFA retain boundary selections.

A again improves DFA, while its primary contrast with FD-DFA spans zero on validation and final test. BP remains stronger. Neither an A advantage nor equivalence to FD-DFA follows (Fig. 9A, FD-study rows). Table 44 reports all four methods, with the primary final-test comparisons collected in Table 45.

Rule	Update seconds	Peak GPU MB	Extra decorrelator state (M)
BP	30.00117 ± 0.00066	111.14	0.00000
DFA	30.00085 ± 0.00059	110.88	0.00000
A	30.00157 ± 0.00123	135.26	0.00000
FD-DFA	30.00175 ± 0.00131	348.92	10.75251

Table 43: Measured work, allocated GPU memory, and additional stored state. Every method has the same supervised parameter count. Decorrelators can be folded into linear maps at inference; their stored state is not a deployment lower bound. MB uses decimal units.

Train	Params (M)	Rule	Test accuracy (%)	Test CE
48k	3.68	BP	63.68 ± 0.33	1.0313 ± 0.0040
48k	3.68	DFA	60.10 ± 0.20	1.1362 ± 0.0043
48k	3.68	A	61.48 ± 0.25	1.1000 ± 0.0058
48k	3.68	FD-DFA	61.56 ± 0.55	1.0938 ± 0.0127

Table 44: Forward-decorrelation test endpoints. Mean $\pm$ sample SD across eight paired seeds at the measured-work endpoint.

Cohort	Primary accuracy contrast	Exact p	Holm p
optimizer	A−DFA	0.0078125	0.0156250
forward	A−FD-DFA	0.7343750	0.7343750

Table 45: Both prespecified final-test accuracy comparisons. Holm adjustment covers this fixed family of two exact sign-flip tests. Secondary results do not replace either primary comparison.

I TARGETED TESTS OF WORK, DAMPING AND NORMALIZATION

The preceding results leave three specific questions: whether activity gains persist with longer measured work, whether error damping should change with width, and how normalization affects predictive stability. We specified these follow-ups before execution, retaining every planned case. They use the same 48,000/2,000 CIFAR-10 training/validation split, paired augmentation, per-example error moments and raw-update norm matching as the shared-BN studies. Training uses an H100 with float32 tensors and TF32 disabled. No test endpoint or intermediate checkpoint selects a setting. All 202 cases finish; 32 are validation-only development runs, and 170 final models each receive one saved test evaluation. Independent float64 log-sum-exp reproduces their recorded prediction metrics. These are follow-up analyses with five paired seeds per condition, not additions to an earlier confirmatory family. Individual t intervals assume approximately normal seed differences and do not provide simultaneous coverage.

I.1 TRANSFER TO LONGER MEASURED WORK

Rule	30 s	60 s	120 s
BP	63.68 ± 0.33	65.82 ± 0.40	67.54 ± 0.19
DFA	60.10 ± 0.20	62.18 ± 0.17	64.67 ± 0.21
A	61.48 ± 0.25	64.21 ± 0.39	66.92 ± 0.30
FD-DFA	61.56 ± 0.55	64.45 ± 0.81	66.88 ± 0.52

Table 46: Test accuracy (%, mean $\pm$ sample SD) at matched measured update work in the 3.68M-parameter CIFAR-10 MLP. The 30-second FD study uses eight paired seeds; the longer budgets share five new paired seeds and transfer the validation-selected settings with separately stretched schedules. They are separate training runs, not checkpoints of one trajectory. Timing excludes validation and export. Appendices H and I give complete cost, selection and endpoint records.

The BP/DFA/A/FD-DFA comparison extends the earlier work budget to 60 and 120 synchronized update seconds. Both budgets reuse the same five new paired seeds (61000–61004) in separate runs. We transfer the previously selected settings unchanged and stretch the same warmup/cosine schedule. This tests transfer across budgets, rather than a newly tuned convergence envelope. The timing definition remains unchanged: sampling, augmentation, conditioning and optimizer work count; validation and export do not. Table 46 gives all absolute test accuracies; the paired contrasts appear in Fig. 9B.

Activity conditioning improves raw DFA at both longer budgets. Its observed gap to BP narrows as the budget increases, although BP remains ahead; forward decorrelation remains competitive. The paired A–FD intervals span zero at both budgets, establishing neither superiority nor equivalence. These results extend the activity gain beyond the original short budget without establishing converged performance or transfer beyond this architecture and dataset.

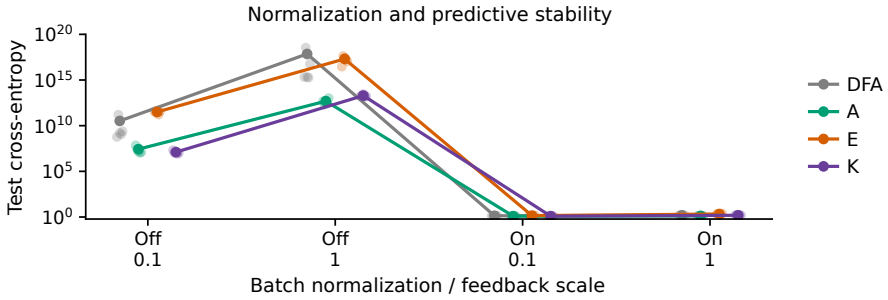

Figure 10: **Normalization restores ordinary predictive-loss scales.** Test cross-entropy across the fixed-recipe normalization/feedback intervention (log scale). Faint points show all five seed endpoints per condition; larger markers and lines show their means. The extreme losses without BN also occur for raw DFA. All conditions and endpoints were specified before execution; there is no test-based selection. The work and width follow-ups appear in Fig. 9B and D.

I.2 Width-specific error damping and a diagonal control

Full and diagonal error conditioning receive the same validation grid: relative damping $\{3, 30, 300, 3000\}$ at each hidden width, evaluated over two development seeds. Final validation CE selects each operator independently; both development runs must be finite, and ties favor larger damping. Full E selects 3000 at both widths, the upper boundary; diagonal E selects 30 and 3, respectively. Five fresh seeds then compare DFA, A, E, diagonal E and K at 5,000 SGD updates, with learning rate 0.1, BN and feedback scale 0.1. A uses relative damping 30; K adds the selected full-E factor without joint search.

Activity gains persist at both widths, while the selected E and diagonal-E rules remain close to DFA, and K remains close to A (Table 47; Fig. 9D). Sparse checkpoint spectra explain what this selected operator does: strong damping makes full E nearly scalar. At the final update its cosine with the raw update exceeds 0.999 in every recorded layer and seed; its damped condition number stays below 1.10. Thus the retuned study largely suppresses the error transformation rather than establishing a robust error-side gain. It neither proves equivalence nor excludes a useful error factor elsewhere.

For an error moment of exact rank r in dimension d, the ridge $\rho\ \mathrm{tr}(C_E)/d$ divided by the mean positive eigenvalue is $\rho r/d$. Fixed relative damping therefore need not preserve conditioning when width changes and rank is batch-limited. The recorded spectra, participation ranks and update angles make this issue observable, but do not establish it as the sole cause of the earlier transfer failure.

	1024–512		2048–1024	
Rule	Accuracy (%)	CE	Accuracy (%)	CE
DFA	51.19 ± 1.97	1.363	47.21 ± 3.19	1.557
A	55.60 ± 0.73	1.253	53.51 ± 2.24	1.353
E	51.34 ± 1.62	1.363	47.32 ± 2.80	1.548
Diagonal E	51.30 ± 1.57	1.365	47.58 ± 3.42	1.527
K	55.58 ± 0.92	1.253	53.71 ± 2.41	1.344

Table 47: Width-specific damping follow-up. Accuracy is mean $\pm$ sample SD across five paired seeds; CE is mean test cross-entropy. The same validation grid selects full and diagonal E independently. K combines A with full E's choice. No test result selects damping.

I.3 Crossing normalization and feedback scale

To isolate two differences between earlier protocols, we cross BN off/on with feedback scale 0.1/1 for DFA, A, E and K, keeping every other setting fixed: hidden widths 1024–512, 5,000 SGD updates, learning rate 0.1 and relative activity/error damping 30. This deliberately fixed recipe is not separately

tuned for each condition. BN restores ordinary predictive-loss scales for all four rules, and activity conditioning retains an accuracy gain within the normalized conditions (Table 48). Reducing feedback scale without BN does not prevent extreme losses. Raw DFA is also unstable, so this behavior is not unique to inverse conditioning.

This intervention identifies a normalization effect under one common recipe; it does not explain every difference between historical protocols. Finite losses are retained regardless of their magnitude. The saved histories report validation loss distributions and logit scales, while sparse checkpoints record update and weight norms. No failing or unfavorable endpoint is replaced.

BN	Feedback	DFA	A	E	K
Off	0.1	24.55 3.31×10^{10}	36.71 2.69×10^{7}	23.93 3.20×10^{11}	38.78 1.28×10^{7}
Off	1	24.80 7.17×10^{17}	38.23 4.85×10^{12}	28.53 2.03×10^{17}	37.77 1.95×10^{13}
On	0.1	51.23 1.36	55.72 1.24	49.33 1.42	56.20 1.24
On	1	46.51 1.55	53.48 1.34	42.06 2.12	50.39 1.55

Table 48: Normalization/feedback intervention. Each cell gives mean test accuracy (%, upper line) and mean CE (lower line) across five paired seeds. All other settings are fixed. Individual endpoints and predictive-loss summaries are retained for the code release.

J PREDICTIVE STABILITY UNDER A SEPARATE SELECTED PROTOCOL

To establish the scope of the shared-BN findings, we retain a separate CIFAR-10 study with different normalization and training choices. It tests whether the ordering of methods persists under that protocol; it does not isolate the effect of any one changed choice.

This comparison uses 45,000 training and 5,000 validation images from a fixed split (seed 15104), training-only channel standardization, and a 3072–1024–512–10 ReLU MLP. Each epoch visits every training identity once with reflected padding of four pixels, random cropping, and horizontal flipping. Batch size is 256, with the partial final batch retained. All methods train for 30 epochs (5,280 updates), using constant-rate SGD without momentum or weight decay and mean softmax CE. A/E/K have no BN and use feedback scale 1.0. These and other differences from the shared-BN studies prevent attributing their different outcomes to one isolated factor.

The ten methods are BP and DFA; each with BN or forward decorrelation; activity-conditioned BP; and A-, E-, K-nDFA. Conditioned hidden gradients match their own raw norms. BN precedes ReLU and freezes running statistics at evaluation. Forward decorrelation uses the same reference mechanism as above, adapted to this study's credit rules. Classifier and bias updates remain supervised/local as appropriate. Training uses float32 with TensorFloat-32 (TF32) disabled and higher-precision reductions as specified in the code.

Selection and numerical failures. Every rule receives twelve initial full-horizon trials and twelve additional distinct trials in a lower-rate grid declared after the first grid's failures. This is adaptive validation development, not a single predeclared joint grid. Raw and BN rules search twelve rates per grid. A and E cross three rates with four dampings; K crosses three rates with four activity/error pairs; decorrelation crosses three rates with four adaptation rates. Thus K's activity damping can differ from A's, and their contrast does not isolate adding E at fixed A. The complete grids are retained for the code release. Final validation accuracy selects the setting, with loss and then case identifier breaking ties; no intermediate or test score enters selection.

Across the completed initial grid and lower-rate grid, 199 trials finish with finite endpoints and 41 are numerical failures. An interrupted first attempt at the initial grid had 92 completed runs, 20 numerical failures, two incomplete cases and six unlaunched cases. The completed rerun resolves the latter eight as numerical failures; it is this rerun, not the interrupted attempt, that enters the development comparison. Several selected settings lie on boundaries. Some conditioned candidates have extreme

finite losses despite competitive accuracy; these remain eligible under the declared accuracy-first rule. Table 50 lists every selected setting and Table 49 retains all final losses.

Five fresh paired replicates follow selection. Development model, feedback, order, and augmentation seeds are 15100–15103. Replicate $r = 0, \ldots, 4$ uses the corresponding seeds $15200 + 10r$ through $15203 + 10r$. All 50 final models are evaluated on the official test set after the complete cohort is fixed. Reported uncertainty is over training seeds conditional on the fixed split and selection procedure.

Interpretation. A does not improve on raw DFA in this protocol, and E performs worse. K exceeds A but uses a different activity damping; it cannot establish the same isolated complementary effect as the main factor studies. Forward-normalized DFA methods outperform the update conditioners. Meanwhile, activity conditioning improves BP, demonstrating that its benefit can also occur with exact gradients. The extreme finite CE values show that norm matching does not guarantee stable predictive losses. These findings qualify the successful shared-BN results without identifying a single cause for their protocol dependence.

Method	Test accuracy (%)	Mean CE	ms/update	Peak MiB
BP	56.83 ± 0.71	1.233	1.84	253
BP + activity	60.27 ± 0.82	1.146	4.17	268
BP + BatchNorm	59.65 ± 0.77	1.138	2.54	254
BP + forward decorrelation	58.35 ± 0.34	1.178	4.16	495
DFA	43.48 ± 0.55	1.634	1.70	253
A-nDFA	41.01 ± 3.42	757.607	3.98	268
E-nDFA	28.93 ± 2.60	1.18×10^{14}	4.01	268
K-nDFA	48.30 ± 2.02	35.140	5.42	268
DFA + BatchNorm	57.15 ± 0.53	1.203	2.51	254
DFA + forward decorrelation	58.33 ± 0.18	1.190	3.94	495

Table 49: **Full-pool CIFAR-10 confirmation.** A 3072–1024–512–10 ReLU MLP trains for 30 epochs (5,280 updates). Each rule receives 24 distinct development trials across the initial grid and a prospectively declared adaptive lower-rate grid and five fresh paired training seeds. Accuracy is mean $\pm$ sample SD, not SEM; CE is mean cross-entropy. Settings are selected by final validation accuracy, with loss breaking ties. Time is the mean synchronized training-step duration on one H200 GPU, including the update and numerical checks; peak memory is the maximum allocated CUDA memory over the five runs. These measurements exclude allocation setup and validation time and are not an equal-wall-clock comparison. Forward decorrelation adapts the reference mechanism to this study's BP/DFA rules.

Method	η	ρ_A	ρ_E	η_{dec}	Val. (%)	Val. CE	Fail/24
BP	0.096	–	–	–	56.80	1.230	6
BP + activity	0.1	1	–	–	61.18	1.137	8
BP + BatchNorm	0.512	–	–	–	58.96	1.162	0
BP + forward decorrelation	0.03	–	–	1e-06	57.82	1.192	12
DFA	0.0005	–	–	–	43.24	1.665	0
A-nDFA	0.01	1	–	–	48.24	429.522	0
E-nDFA	0.03	–	10	–	33.38	1.50×10^{14}	0
K-nDFA	0.01	3	3	–	52.62	18.424	0
DFA + BatchNorm	0.016	–	–	–	56.72	1.221	0
DFA + forward decorrelation	0.003	–	–	1e-05	58.00	1.193	15

Table 50: **Complete development selection.** Final validation accuracy selects the configuration; ties use lower cross-entropy and then case ID. A dash denotes an inactive parameter. Every numerical failure counts in the 24-trial budget. Relative damping is applied separately as $\max(\rho\ \mathrm{mean}(X^2), 10^{-6})$. Values are from the single development seed; they are not confirmation test estimates.

Independent predictive-loss audit. To distinguish rare extreme errors from broad miscalibration, we recomputed per-example CE from every saved test logit using float64 log-sum-exp. All 50 endpoint means agree with the original evaluation. Table 52 reports seed-averaged distribution

Contrast	Mean (pp)	Individual 95% CI	Holm p_t	Exact p
A-nDFA − DFA	-2.47	[-6.96, +2.01]	0.2002	0.2500
E-nDFA − DFA	-14.55	[-18.16, -10.94]	0.0018	0.0625
K-nDFA − A-nDFA	+7.29	[+2.75, +11.84]	0.0224	0.0625
A-nDFA − DFA + forward decorrelation	-17.32	[-21.67, -12.98]	0.0018	0.0625
A-nDFA − DFA + BatchNorm	-16.14	[-20.93, -11.36]	0.0022	0.0625
BP + activity − BP	+3.44	[+2.82, +4.07]	0.0006	0.0625

Table 51: **All six prespecified paired comparisons.** Intervals and t tests use five paired differences (four degrees of freedom) and an approximate-normality assumption that is difficult to assess at this sample size. Holm correction covers the six t tests; intervals are individual, not simultaneous. The separate two-sided sign-flip test enumerates all 32 assignments and assumes sign exchangeability. Its minimum attainable p is 0.0625 for five nonzero pairs. Constant differences give undefined t statistics and intervals, which are omitted rather than reported as certainty.

summaries for the raw and conditioned DFA rules. The large median losses for A and E show that their instability extends well beyond a small tail. The highest-loss one percent of examples account for a minority of the total loss. This confirms a predictive failure without establishing that any one implementation choice caused it.

Rule	Median CE	99th-percentile CE	Top 1% loss share (%)
DFA	1.38	5.78	4.3
A	119.44	7179.52	12.5
E	4.81×10^{13}	8.52×10^{14}	8.3
K	2.69	273.07	9.7

Table 52: Loss distributions in the separate full-pool protocol. Each entry averages the indicated per-seed statistic over all five frozen test endpoints; quantiles are not computed from pooled predictions. The top-one-percent share is the fraction of total CE contributed by those examples.

K Moment orientation and imperfect credit

This experiment separates two questions: how a local moment changes an approximate update, and whether that change improves its relation to the true gradient. We construct finite samples with prescribed activity and error moments while holding their raw cross moment fixed. The targets use the desired true gradient explicitly. This is an engineered diagnostic of the operator, not a local procedure for discovering useful credit or unknown noise.

Constructing the moments within a fixed-feedback network. The prescribed moments must be compatible with a single network, rather than assigned to unrelated matrices. The construction below ensures this while allowing activity and error orientations to vary independently.

Let $A \in \mathbb{R}^{N\times 8}$ contain activities and $D \in \mathbb{R}^{N\times 8}$ contain local DFA teaching signals, with $N \geq 16$. In this subsection samples are rows, so the feedback matrix B is transposed relative to Eq. 1. For orthogonal bases $Q_L, Q_R \in \mathbb{R}^{8\times 8}$, set

$$U = Q_L Q_R^\top/\sqrt{8}, \qquad g = Q_L \operatorname{diag}(1,1,1,1,0,0,0,0) Q_R^\top/2.$$

Both matrices have Frobenius norm one, U has full rank, and $\cos(g,U) = 1/\sqrt{2}$. Assign $C_A = Q_R \operatorname{diag}(a) Q_R^\top$ and $C_E = Q_L \operatorname{diag}(e) Q_L^\top$ independently. Each eigenvalue vector is either isotropic $(8.5,\ldots,8.5)$, signal-low $(1,1,1,1,16,16,16,16)$, or signal-high with these halves exchanged. The two anisotropic states have exactly the same spectrum; the isotropic control preserves their trace. The first four coordinates contain the task gradient in all orientation cells.

Choose an orthonormal $N \times 16$ matrix $[Q_A, Q_Z]$ and construct

$$A = \sqrt{N} Q_A \operatorname{diag}(\sqrt{a}) Q_R^\top, \qquad D = A C_A^{-1} U^\top + \sqrt{N} Q_Z R^\top, \quad RR^\top = C_E - U C_A^{-1} U^\top.$$

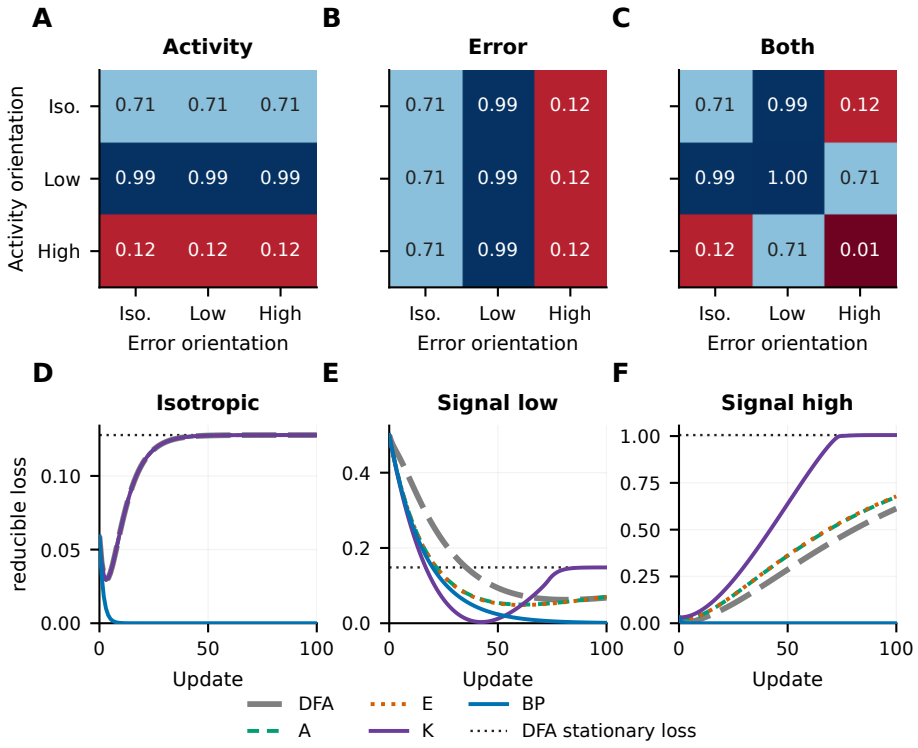

Figure 11: **Conditioning changes direction without repairing credit. A–C,** True-gradient cosines for A, E and K with independently assigned moment orientations and damping $1/1$; the raw cosine is 0.707 throughout. Colors use the common cosine range $[0, 1]$; labels round to two decimals. The displayed 1.00 is 0.999904, not exact credit recovery. Iso., low and high denote isotropic, signal-low and signal-high moments. **D–F,** Prescribed 100-update trajectories with both moments isotropic, signal-low, or signal-high. All methods start from the same weights within each cell; the loss shown excludes the constant readout-complement component. Initial direction improvement can accelerate approach to a suboptimal DFA stationary point in this diagnostic with a fixed bottleneck readout. With isotropic moments, the four local trajectories coincide and finish above their initial loss. These are engineered targets and a fixed step size, with all moment/damping cells available in the reproduction records.

The residual covariance is positive definite. Direct multiplication gives $A^\top A/N = C_A$, $D^\top D/N = C_E$, and $D^\top A/N = U$. Changing the error moment changes a teaching component orthogonal to activity, leaving the raw update unchanged. Activity interventions use the corresponding compensating construction of D.

The trainable matrix is $W \in \mathbb{R}^{8\times 8}$, initialized at zero. A fixed bottleneck readout and its orthogonal complement are $V = [I; I]/\sqrt{2}$ and $V_\perp = [I; -I]/\sqrt{2}$. Use the same fixed feedback $B = \frac{1}{2}V + \sqrt{3/4}V_\perp$ in every condition. With $D_\star = AC_A^{-1}g^\top$, define the initial output residual E_0 and squared loss by

$$E_0 = D_\star V^\top + (D - \tfrac{1}{2}D_\star)V_\perp^\top/\sqrt{3/4}, \qquad L(W) = \frac{\|E_0 + AW^\top V^\top\|_F^2}{2N}.$$

Then $E_0 V = D_\star$ and $E_0 B = D$: the actual BP and DFA weight gradients at zero are precisely g and U. The readout and feedback lie in $\mathbb{R}^{16\times 8}$ and have full column rank eight, with $V^\top V = B^\top B = I$ and $V^\top B = I/2$ throughout. The exact Hessian acts as $\Delta W \mapsto \Delta W C_A$; its left factor is I, independently of the assigned C_E. Any error-side benefit here therefore concerns the direction of imperfect credit, rather than identification of a Hessian factor.

Predicted direction changes. Because the raw update and true gradient are fixed at initialization, any change in their alignment comes from conditioning. The shared eigenvector bases make that change explicit.

Raw, A-, E-, and K-nDFA share each cell's samples and targets. Each conditioned matrix is matched to $\|U\|_F$. For independently crossed $\lambda_A, \lambda_E \in \{0.1, 1, 16\}$, let w_i be the product of the active reciprocal factors $(a_i + \lambda_A)^{-1}$ and $(e_i + \lambda_E)^{-1}$. The predicted true-gradient cosine is

$$\cos(g, \widetilde{U}) = \frac{\sum_{i=1}^{4} w_i}{\sqrt{4 \sum_{i=1}^{8} w_i^2}}.$$

At equal damping, a signal-low factor improves the direction, while exchanging its eigenspaces worsens it. Two favorable factors reinforce one another; opposing orientations cancel at equal damping. Thus the spectrum alone does not determine the sign of a benefit, even with exact moments (Fig. 11A–C).

Eight orthogonal basis seeds and $N \in \{32, 64, 128\}$ verify the construction across all nine orientation pairs and nine damping pairs. These realizations check equivariance and moment identities; they are not independent tasks. All conditions and direct numerical checks are retained for the code release.

Credit quality and the rotation boundary. An update that points partly along the true gradient can still be turned away from it by conditioning. We now bound how large a rotation is safe, given a lower bound on the raw update's alignment.

For vectorized $u \neq 0$, a symmetric positive-definite (SPD) conditioner P, and its positively norm-matched direction $v \propto Pu$, let $\theta = \angle(u, v) \in [0, \pi/2)$ and assume $\cos(g, u) \geq c > 0$. Decomposing the unit vectors parallel and perpendicular to u gives the sharp cone bound

$$\cos(g, v) \geq c \cos\theta - \sqrt{1 - c^2} \sin\theta. \tag{26}$$

The lower bound is attained by placing the perpendicular component of g opposite that of v. Every gradient in this cone therefore gives descent along $-v$ exactly when $c > \sin\theta$; equality allows zero derivative. This is an elementary application of classical operator-angle geometry (Gustafson, 2010), not a new angle theorem. For an SPD factor P, $\sin\angle(u, Pu) \leq (\kappa(P) - 1)/(\kappa(P) + 1)$. The vectorized K-nDFA factor has the product of the two factor condition numbers. A true-gradient margin is an additional assumption: local covariance estimates do not certify it.

We check the boundary by fixing the signal-low/signal-low construction and varying the target gradient within this cone. Choosing its perpendicular component opposite to the conditioned update gives ascent, zero derivative and descent below, at and above the boundary, while raw DFA remains a descent direction. Direct loss calculations also verify

$$L(-\eta\widetilde{U}) - L(0) = -\eta\langle g, \widetilde{U}\rangle + \frac{\eta^2}{2} \operatorname{tr}(\widetilde{U} C_A \widetilde{U}^\top).$$

Thus a directional guarantee alone does not certify an arbitrary finite step: at zero derivative, the quadratic term raises the loss.

Short trajectories and preserved stationary mismatch. The one-step calculation leaves open what happens over repeated updates. We therefore follow each constructed network from the same initial weights within each condition. Figure 11D–F displays the isotropic, signal-low, and signal-high cases. In the isotropic case the local-rule trajectories coincide; the anisotropic cases show that favorable initial alignment need not yield a better endpoint.

All 81 orientation/damping cells receive 100 full-batch updates at the predeclared learning rate 0.03, with seed 9151500 and $N = 64$. Raw DFA, A, E, K, and BP give 405 trajectories. A/E/K recompute the current error moment and match their current raw-DFA norm at every step; BP uses its exact gradient. There are no numerical failures. This is one diagnostic step size, without optimizer tuning or selection of intermediate checkpoints.

The true and approximate gradients obey

$$g(W) = g(0) + WC_A, \qquad U(W) = U(0) + \tfrac{1}{2} WC_A.$$

Their stationary weights differ: $W_{\mathrm{BP}} = -g(0)C_A^{-1}$ and $W_{\mathrm{DFA}} = -2U(0)C_A^{-1}$. Invertible conditioning preserves the zero set of the raw update and cannot remove this mismatch. The loss in the readout complement remains constant; we compare the reducible component $\|E(W)V\|_F^2/(2N)$ within each cell. Initial absolute losses are not comparable measures of difficulty across the constructed targets.

Scope of the stationary mismatch. The mismatch depends on information lost through the fixed readout. To make that restriction explicit, we compare this construction with a linear network whose feedback and readout preserve all output directions.

Full column rank here does not make credit injective in the 16-dimensional output space: $V^\top$ and $B^\top$ each have an eight-dimensional kernel. The fixed readout cannot change $E(W)V_\perp = E_0V_\perp$, whereas this residual contributes to DFA because $V_\perp^\top B = \sqrt{3/4}I$. The mismatch therefore relies on a fixed bottleneck readout and its unlearnable output complement. For comparison, consider a linear fixed-readout model with activity dimension d, hidden width h, and $m \le h$ outputs. Define the residual–activity cross moment $R(W) = E(W)^\top A/N \in \mathbb{R}^{m\times d}$. If both $V^\top, B^\top \in \mathbb{R}^{h\times m}$ have full column rank m, then

$$g(W) = V^\top R(W) = 0 \quad \Longleftrightarrow \quad R(W) = 0 \quad \Longleftrightarrow \quad U(W) = B^\top R(W) = 0.$$

This condition concerns the cross moment, not necessarily zero sample residual. Thus the present stationary mismatch is not a general property of full-rank linear DFA. Neither this counterexample nor the injectivity argument establishes the stationary structure of nonlinear networks with activation-dependent gates or trained readouts; their learning behavior requires separate experiments.

With both signal-low factors, K initially points closest to the true gradient among the local rules, but approaches the mismatched stationary point faster and has a higher final reducible loss in the displayed trajectory. An improved initial direction therefore need not imply a better endpoint (Fig. 11E). The complete orientation/damping grid remains in the reproduction records.

L Connections to other credit-assignment rules

Conditioning changes how a credit signal is used while retaining DFA's fixed random feedback. It is complementary to methods that change the source or routing of credit.

Feedback pathways. FA/DFA performance depends on architecture and task (Lillicrap et al., 2016; Nøkland, 2016; Launay et al., 2020; Bartunov et al., 2018). The align-then-memorize analysis concerns acquisition of useful feedback (Refinetti et al., 2021), whereas our spectral calculation assumes alignment. Weight mirrors and feedback matching adapt feedback directions (Akrout et al., 2019; Lee and Lee, 2026; Moskovitz et al., 2018). Low-rank parameterizations and orthogonalization address related limitations (Roy et al., 2026; Boeshertz et al., 2026). Invertible conditioning preserves algebraic rank but changes singular values and effective rank.

Alternative credit signals. Local objectives, target projection, predictive coding, target propagation and equilibrium propagation change how hidden layers receive credit (Nøkland and Eidnes, 2019; Frenkel et al., 2021; Whittington and Bogacz, 2017; Lee et al., 2015; Scellier and Bengio, 2017). Perturbation rules estimate credit from perturbation–error correlations (Williams, 1992; Fiete et al., 2007); NMNC restricts perturbations to an activity manifold (Kang et al., 2026) and is competitive in the exploratory Fashion-MNIST sweep. Conditioning does not resolve every limitation of approximate credit.

Curvature and locality. Natural-gradient and Kronecker-factored methods motivate the matrix form (Amari, 1998; Martens and Grosse, 2015; Grosse and Martens, 2016; George et al., 2018). Forward decorrelation changes representations (Desjardins et al., 2015; Huang et al., 2018; 2019) and can help perturbation learning and BP (Dalm et al., 2026; 2024; Ahmad, 2024). nDFA instead conditions updates. Our implementation requires layer-wide statistics and solves; synaptic locality remains an open implementation question.