ISOTROPIC UPDATES ENABLE BACKPROP-FREE TRAINING OF MODERN VISION ARCHITECTURES

Anonymous authors
Paper under double-blind review

ABSTRACT

Forward Direct Feedback Alignment (FDFA) trains networks without backpropagation (BP) by delivering the output error directly to each layer through feedback matrices learned with forward-mode automatic differentiation. On modern vision architectures, namely deep non-convolutional models with normalization and residual connections such as ViT, MLP-Mixer, and Swin, however, FDFA and existing backpropagation-free methods fail to train well. We hypothesize that a key contributing factor is spectral concentration of the updates: the errors of all samples pass through the same feedback matrix, and in our diagnostics FDFA updates are skewed toward a few singular directions. Accumulated skewed updates inflate the weights anisotropically and lower the sensitivity of some feedback points, where single-pass feedback estimation then degrades and in turn sustains the skewed updates. We therefore introduce norm-preserving update orthogonalization (NPO), which makes the update of any optimizer isotropic while preserving its column and row spaces and its Frobenius norm, and apply it to FDFA to obtain IsoFDFA. Training diagnostics are consistent with the hypothesis, and NPO suppresses each stage. IsoFDFA trains MLP-Mixer, ViT, and Swin on CIFAR, matching or exceeding a BP baseline trained with the same learning rate on MLP-Mixer and ViT, and trains DeiT-S on ImageNet, where FDFA makes no progress. To our knowledge, this is the first result showing that a vision Transformer can be trained on ImageNet to non-trivial accuracy with learned feedback and a single forward-mode pass per step, without auxiliary local losses or backpropagation of errors across blocks.

1 INTRODUCTION

The success of deep learning rests on the efficient gradient computation of backpropagation (BP) (Rumelhart et al., 1986). However, BP propagates the error toward the input through the transpose of the forward weights, which requires the forward and backward weights to correspond (weight transport) (Grossberg, 1987; Crick, 1989), and the gradient of each layer must wait until the error has propagated from the output side (update locking) (Jaderberg et al., 2017). These properties are an issue for the biological plausibility of BP (Lillicrap et al., 2020) and also constrain the design of hardware in which layers learn locally and in parallel.

Direct Feedback Alignment (DFA) (Nøkland, 2016) sends the output error through a fixed random matrix directly to the output of each block; we call these positions *feedback points*. Once the output error is available, each layer can compute a local gradient without waiting for sequential backpropagation from downstream, and no symmetry between the forward and feedback weights is required. Training Transformers with DFA has been reported (Launay et al., 2020), but a gap to BP remains, and its size depends on the granularity of the feedback and on the optimization settings. To narrow this gap, one line of work improves the agreement between the feedback and the true Jacobian or gradient.

This line of work learns the feedback from forward-mode derivative information, as in Forward Direct Feedback Alignment (FDFA) (Bacho & Chu, 2024) and GrAPE (Caillon et al., 2026). FDFA tracks the mean Jacobian from each feedback point to the output, whereas GrAPE uses a cosine-alignment objective and, in its full version, periodic BP anchor steps. We focus on single-pass FDFA, which estimates the feedback at all points from one combined forward-mode pass. In our experi-

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

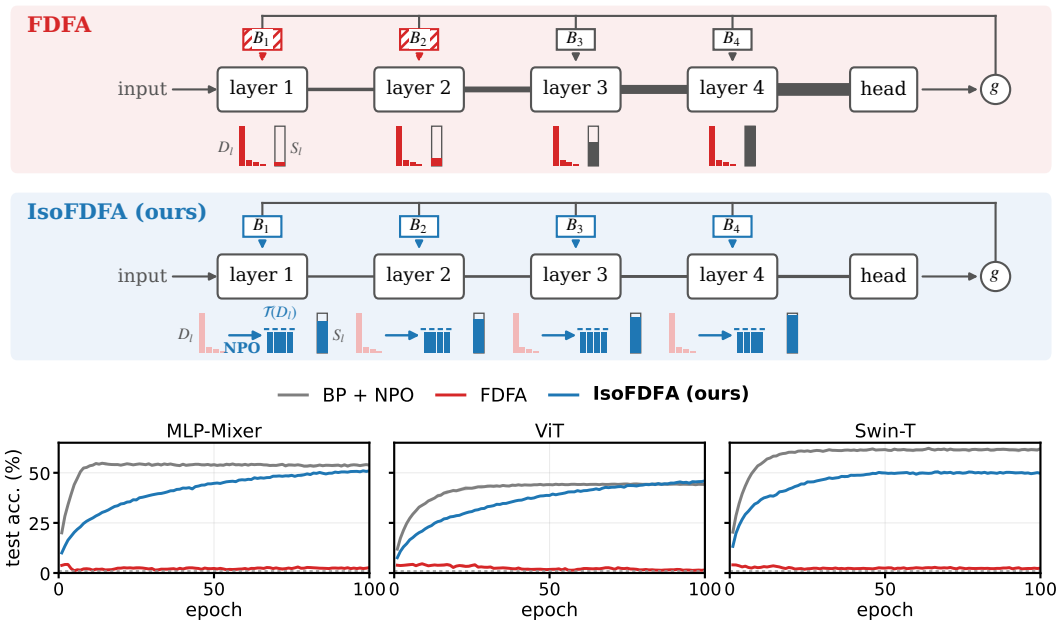

Figure 1: Failure of FDFA and its fix by IsoFDFA. **(Top)** Line width between blocks indicates the activation norm; bars below each block show the singular values of the raw update D_l and the sensitivity S_l. In FDFA, the raw updates are *skewed*, with their energy concentrated on a few singular directions; they inflate the activations, the sensitivity drops, and their learned feedback matrices fail to align with the Jacobian (hatched). IsoFDFA applies NPO to the same raw update and makes it *isotropic*, equalizing its singular values (dashed line) while preserving its norm; the sensitivities and feedback matrices recover (blue). **(Bottom)** Test accuracy on CIFAR-100 (dotted: chance level); BP + NPO is shown as the BP reference for a fair comparison (Section 4.3). Lines show the mean over three seeds, and shaded regions indicate ± 1 standard deviation; some bands are too narrow to be visible at this scale.

ments, this method struggles to train modern vision architectures, namely deep non-convolutional models with normalization and residual connections such as ViT (Dosovitskiy et al., 2021), MLP-Mixer (Tolstikhin et al., 2021), and Swin (Liu et al., 2021).

We hypothesize that spectral concentration in the updates contributes to this failure through its interaction with feedback estimation. In FDFA, errors delivered to a block pass through a feedback matrix shared across samples and therefore lie in a common subspace. In our experiments, the resulting optimizer updates are *skewed*: they exhibit low stable rank $\|D\|_F^2/\|D\|_2^2$, with most of their energy concentrated in a few singular directions. If dominant directions persist across steps, such skewed updates can accumulate into anisotropic, enlarged weights.

We further hypothesize that the associated activation-scale changes create imbalances across feedback points in their sensitivity, the expected squared norm of the Jacobian of the output with respect to the point. In single-pass estimation, the directional derivative combines the contributions of all points, and reduced sensitivity at one point relative to the others increases the relative contribution of the other points, which act as cross-point noise, to its Jacobian estimate. This places a greater burden on feedback estimation, while degraded feedback can sustain updates that fail to restore balanced sensitivities. Training diagnostics and interventions on the update spectrum support this hypothesized interaction.

Motivated by this hypothesis, we introduce *norm-preserving update orthogonalization* (NPO). NPO replaces the raw optimizer update with its polar factor rescaled to the original Frobenius norm. The exact transformation makes the update *isotropic*, equalizing its nonzero singular values while preserving the column and row spaces and the total update magnitude. This reduces the spectral norm of the update, its maximal action on a unit input, under these constraints without introducing a separate step-size scaling rule. NPO leaves feedback estimation and delivery unchanged and can be applied to different optimizers. We call FDFA equipped with NPO *IsoFDFA*. The polar factor itself

is standard in orthogonalized optimizers such as Muon; we use it in a norm-preserving form to study and mitigate the interaction between update spectra and learned feedback in single-pass FDFA.

Our contributions are as follows.

- We propose a failure mechanism in which spectral concentration in the updates interacts with learned feedback in single-pass FDFA (Section 2.2). Training diagnostics support this mechanism.
- We introduce NPO and apply it to FDFA to obtain IsoFDFA. IsoFDFA trains ViT, MLP-Mixer, and Swin (Liu et al., 2021) on CIFAR, where FDFA stays near chance level on CIFAR-100, and reaches accuracy comparable to our BP baseline on MLP-Mixer and ViT. It also trains DeiT (Touvron et al., 2021) on ImageNet with a single forward-mode pass per step.

2 METHOD

2.1 PRELIMINARIES: FDFA

We divide the network into L blocks, $h_l = f_l(h_{l-1}; W_l)$, followed by a classification head $o = f_{\text{head}}(h_L; W_{\text{head}}) \in \mathbb{R}^C$, where C is the number of classes. Here $h_l \in \mathbb{R}^{n_l}$ is the flattened block output. The block outputs at which the error is delivered are called *feedback points*; there are L of them, excluding the head. For a mini-batch of N samples, let $g_i = \partial \ell_i / \partial o_i$ and $J_{l,i} = \partial o_i / \partial h_{l,i}$. Direct feedback alignment (DFA) (Nøkland, 2016) replaces the exact block-output error $J_{l,i}^\top g_i$ with feedback through a fixed matrix $B_l \in \mathbb{R}^{C \times n_l}$:

$$\delta_{l,i} = B_l^\top g_i, \qquad G_l = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{\partial h_{l,i}}{\partial W_l} \right)^\top \delta_{l,i}. \tag{1}$$

Local differentiation treats the block input and delivered error as constants. Every block receives feedback, while the head is updated directly from g_i. This method is regarded as *backpropagation-free* because the error reaches each block directly from the output, without being propagated sequentially backward across blocks; differentiation within each block is still performed. Throughout, we use the term in this sense.

FDFA (Bacho & Chu, 2024) learns B_l to track the mean Jacobian $\bar{J}_l = \mathbb{E}_x[J_l(x)]$ using forward-mode automatic differentiation. Independent, unnormalized probes $u_{l,i} \sim \mathcal{N}(0, I_{n_l})$ are injected at every feedback point. One combined forward-mode pass yields

$$d_i = \sum_{m=1}^{L} J_{m,i} u_{m,i}, \qquad \widehat{J}_l = \frac{1}{N} \sum_{i=1}^{N} d_i u_{l,i}^\top. \tag{2}$$

For a fixed model, averaging over probes recovers the batch-mean Jacobian, and averaging over data recovers $\bar{J}_l$. FDFA updates B_l at every step so as to decrease $\frac{1}{2}\|B_l - \widehat{J}_l\|_F^2$, which under plain gradient descent is an exponential moving average of $\widehat{J}_l$. Per-point estimation probes each point separately and requires L forward-mode passes; *single-pass* FDFA, the original formulation of Bacho & Chu (2024) and the one we study, uses one.

2.2 FAILURE MECHANISM HYPOTHESIS

We hypothesize that spectral concentration in FDFA updates can destabilize both the forward network and its learned feedback.

Shared feedback and concentrated updates. Under BP, the error $J_{l,i}^\top g_i$ of each sample passes through a sample-specific Jacobian, and the update can therefore spread over directions that differ from sample to sample. Under FDFA, in contrast, the errors of all samples pass through the single shared matrix B_l and therefore lie in the common subspace $\mathrm{Im}(B_l^\top)$. For example, for a linear block $h_l = W_l h_{l-1}$ that receives feedback directly at its output, $G_l = B_l^\top \left(N^{-1} \sum_i g_i h_{l-1,i}^\top \right)$, and therefore $\mathrm{col}(G_l) \subseteq \mathrm{Im}(B_l^\top)$ and $\mathrm{rank}(G_l) \leq C$. This relation is specific to linear blocks and

does not determine the degree of skew. In our training diagnostics, however, the actual optimizer updates exhibit strong spectral concentration, as measured by the stable rank $\|D\|_F^2/\|D\|_2^2$, which equals the rank when the nonzero singular values are equal and approaches 1 when one dominates: for ViT, their median stable rank stays between 1 and 2 under FDFA, compared with 3–8 under BP (Figure 2a), with most of the update energy in a few singular directions. We refer to such updates as *skewed*.

Weight growth and sensitivity imbalance. For the same Frobenius norm, a skewed update acts more strongly along its dominant input directions. When successive updates reinforce the same directions, they accumulate into large, anisotropic changes in the weights, as derived in Appendix A. In networks with normalization and residual connections, changes in activation scale alter downstream Jacobians. We hypothesize that weight and activation growth reduces the sensitivity $S_l = \mathbb{E}_x\|J_l(x)\|_F^2$ of some feedback points relative to others, widening the sensitivity imbalance. In our ViT diagnostics, FDFA exhibits increasing weight anisotropy and growing weight and activation norms, while the ratio of the largest to the smallest sensitivity rises from 10^2 to 10^9 (Figure 2b–c).

Consequences for feedback estimation. Together with the sensitivity, we quantify the relative contamination from the other points, the *cross-point noise*, by

$$S_l = \mathbb{E}_x\|J_l(x)\|_F^2, \qquad \chi_l = \frac{\sum_{m\neq l} S_m}{S_l}. \tag{3}$$

For independent unit-covariance probes, S_l is the expected energy of the contribution $J_l u_l$ to the directional derivative, while $\sum_{m\neq l} S_m$ is that of the combined contribution from the other points; χ_l is therefore the ratio of the expected energy contributed by the other points to that of point l. For the variance analysis, fix the model and a mini-batch $\mathcal{B}$, and consider unpooled feedback, i.e., B_l acting on the full flattened block output, without cross-sample coupling. Let $S_l^{\mathcal{B}} = N^{-1}\sum_i \|J_{l,i}\|_F^2$ and $\chi_l^{\mathcal{B}} = \sum_{m\neq l} S_m^{\mathcal{B}}/S_l^{\mathcal{B}}$. For independent Gaussian probes and $S_l^{\mathcal{B}} > 0$,

$$\frac{\mathbb{E}_u\big\|\widehat{J}_l - \mathbb{E}_u[\widehat{J}_l]\big\|_F^2}{S_l^{\mathcal{B}}} = \frac{n_l+1}{N} + \frac{n_l\chi_l^{\mathcal{B}}}{N}, \tag{4}$$

which is derived in Appendix A. The first term comes from the point's own probe and the second from the other points; per-point estimation removes only the second. The estimator remains unbiased, but at a fixed point the normalized variance grows linearly with $\chi_l^{\mathcal{B}}$, and a point whose sensitivity falls relative to the others is increasingly contaminated. The learned B_l averages these single-step estimates across steps but must also track a drifting target, and we expect the contamination to degrade it at low-sensitivity points. We hypothesize that this closes a vicious cycle: degraded feedback yields updates that are again skewed and fail to restore balanced sensitivities, which further degrades the feedback.

2.3 Norm-preserving update orthogonalization

To retain the contribution of the learned feedback while reducing spectral concentration, we introduce *norm-preserving update orthogonalization* (NPO). NPO acts on the raw update that the optimizer produces from the errors delivered through the learned feedback matrix B_l. Consider one selected weight matrix W_t of block l at step t, and let $D_t = W_t^{\mathrm{raw}} - W_t$ be its raw optimizer displacement at step t; where the block matters, as in Figure 1, we write it D_l. For nonzero $D_t = U_t\Sigma_t V_t^\top$ with compact SVD and rank r_t, define

$$\mathcal{T}(D_t) = \frac{\|D_t\|_F}{\sqrt{r_t}} U_t V_t^\top, \qquad W_{t+1} = W_t + \mathcal{T}(D_t), \tag{5}$$

with $\mathcal{T}(0) = 0$; $U_t V_t^\top$ is the orthogonal polar factor of D_t. We call a nonzero update *isotropic* if its nonzero singular values are equal, i.e., it acts with the same gain in every direction of its row space. An isotropic update is the opposite extreme of the skewed updates of Section 2.2: its stable rank equals its rank, whereas that of a skewed update is close to one. NPO produces an isotropic update while keeping the column and row spaces and the Frobenius norm of the raw displacement, which are determined by the learned feedback and the optimizer; only the distribution of energy across the singular directions changes. The result discards the relative magnitudes of the singular values but

remains positively aligned with the raw update, $\cos(D_t, \mathcal{T}(D_t)) = \|D_t\|_*/(\sqrt{r_t}\,\|D_t\|_F) \geq 1/\sqrt{r_t}$, where $\|\cdot\|_*$ is the nuclear norm.

Among matrices with the same column and row spaces and Frobenius norm as D_t, Eq. (5) minimizes the spectral norm; among these minimizers, it is closest to D_t in Frobenius distance (Fan & Hoffman, 1955; Higham, 1989); Appendix A gives the derivation. NPO therefore reduces the largest action of an update without changing its total magnitude, with the scale inherited from the optimizer. Section 3 discusses its relation to existing polar-based updates.

For $D_t \neq 0$, we compute an approximate polar factor Q_t using eight Newton–Schulz iterations with the Muon coefficients (Jordan et al., 2024), and apply $W_{t+1} = W_t + \|D_t\|_F\, Q_t/\|Q_t\|_F$. The explicit rescaling preserves the update norm up to numerical precision. The finite Newton–Schulz iteration only approximately equalizes the singular values; the spectral-norm optimality above therefore need not hold exactly. The transformed set $\mathcal{W}_{\mathrm{NPO}}$ contains two-dimensional and reshaped four-dimensional forward-model parameters, excluding the classifier and position-dependent bias tables such as those of Swin; positional embeddings and class tokens are not included. Detailed selection and reshaping rules are in Appendix B. Other parameters retain their optimizer updates. Optimizer states, feedback estimation, and feedback delivery are unchanged. Any weight decay included in the raw displacement is transformed jointly with the remaining update. We call FDFA with this operation *IsoFDFA*. NPO acts independently on each matrix and adds no forward-mode passes; Algorithm 1 gives the complete procedure.

3 Related Work

DFA and FDFA. Feedback Alignment showed that fixed random feedback weights suffice for learning because the forward weights align to them (Lillicrap et al., 2016), and DFA sends the output error directly to every layer, removing the sequential backward pass (Nøkland, 2016). The dynamics of alignment have been analyzed (Refinetti et al., 2021), but the accuracy of DFA falls short of BP: it fails on CNNs (Bartunov et al., 2018; Launay et al., 2019) and leaves a gap on Transformers (Launay et al., 2020), although with fixed random feedback it has been scaled to billion-parameter Transformers and Vision Transformers on optical hardware (Wang et al., 2026). FDFA (Bacho & Chu, 2024) and GrAPE (Caillon et al., 2026) improve the accuracy of the feedback by learning the feedback matrices from unbiased forward-gradient estimates (Baydin et al., 2022; Ren et al., 2023); GrAPE additionally inserts periodic BP anchor steps. Boeshertz et al. (2026) identify low-dimensional updates and trajectories in FA and mitigate them with Muon and activity normalization. We instead study single-pass FDFA, where the feedback is learned and interacts with the update spectrum, and we keep the Frobenius norm of the raw update rather than adopting the step size of Muon; this combination substantially improves single-pass FDFA.

Orthogonalized updates and spectral optimizers. Orthogonalizing the gradient to speed up training was proposed by Tuddenham et al. (2022), and Muon (Jordan et al., 2024), which orthogonalizes the momentum of the gradient by Newton–Schulz iterations, is now widely used. Bernstein & Newhouse (2024) interpret such updates as steepest descent under the spectral norm, and PolarGrad (Lau et al., 2025) unifies matrix-gradient optimizers from a preconditioning perspective. PolarAdamW (Zhang, 2026) applies a polar map to the preconditioned AdamW direction and is algorithmically the closest to our rule. Our rule differs from Muon in what is orthogonalized: Muon orthogonalizes the momentum of the gradient, whereas we orthogonalize the raw update of an arbitrary optimizer, so that Adam's per-coordinate scaling shapes the update before it is orthogonalized. It differs from both Muon and PolarAdamW in how the result is scaled: they use a learning rate and a shape-dependent factor, whereas we restore the Frobenius norm of the raw update, leaving the step size of every block to the optimizer. In our experiments, simply replacing the optimizer of FDFA by Muon does not suffice (Section 4.4). Orthogonality regularization of the weights (Bansal et al., 2018) acts on the weights rather than on the updates; it is not a component of our method and is treated as a separate control.

4 EXPERIMENTS

We examine whether the failure mechanism described in Section 2.2 actually occurs in FDFA and is removed by IsoFDFA (Section 4.2), and whether IsoFDFA enables backpropagation-free training of modern vision architectures (Section 4.3). We then identify which components of the method matter through ablations (Section 4.4).

4.1 EXPERIMENTAL SETUP

Datasets and models. We use CIFAR-10 and CIFAR-100 (Krizhevsky & Hinton, 2009) with ResNet-18 (He et al., 2016), MLP-Mixer (Tolstikhin et al., 2021), a small ViT (Dosovitskiy et al., 2021), and Swin-T (Liu et al., 2021), and ImageNet-1k (Deng et al., 2009) with DeiT-S/16 (Touvron et al., 2021). The ViT follows the configuration of Flügel et al. (2025), with 6 blocks, 4 heads, width 256, MLP width 512, and patch size 16 (about 3.4M parameters); Swin-T and DeiT-S/16 are the standard configurations. All models take 224×224 inputs; on CIFAR, the only augmentations are random resized crops to 224×224 and horizontal flips. We report test accuracy. We refer to them simply as ViT, Swin, and DeiT below. ResNet-18 serves as a convolutional control on which FDFA does not fail. Further details are given in Appendix C.1.

Baselines. On CIFAR, we compare IsoFDFA with BP and with two families of backpropagation-free methods. The first is the DFA family: DFA (Nøkland, 2016), FDFA (Bacho & Chu, 2024), and GrAPE (Caillon et al., 2026) without its backpropagation anchor steps. FDFA, GrAPE, and IsoFDFA use single-pass estimation (Section 2.1). The number of feedback points L equals the number of blocks excluding the classification head: 8 for ResNet-18, 8 for MLP-Mixer, 6 for ViT, 12 for Swin, and 12 for DeiT (Appendix C.2). We compare with per-point FDFA, which uses L passes, in the ablations. The second is the forward-gradient family, which trains the forward weights directly from forward-mode directional derivatives without any feedback matrix: forward gradient (FG) (Baydin et al., 2022) and multi-tangent forward gradient (MTFG) (Flügel et al., 2025) with $k = 4$ tangents. On ImageNet, we compare BP, FDFA, and IsoFDFA.

Training. On CIFAR, all methods use Adam (Kingma & Ba, 2015) at a constant learning rate of 10^{-4} for 100 epochs. The same learning rate is used for all methods and models, and the feedback matrices of FDFA, GrAPE, and IsoFDFA are initialized to zero and updated with Adam at learning rate 10^{-4} rather than with the plain EMA. On ImageNet, we follow the DeiT recipe (Touvron et al., 2021) with AdamW (Loshchilov & Hutter, 2019) for 300 epochs, at learning rate 10^{-3} for BP; for FDFA and IsoFDFA, we use the same settings, with learning rate 2.5×10^{-4}, weight decay also on LayerNorm, and feedback matrices updated with AdamW with 5 warm-up epochs (Table 5). There, NPO is applied to the full AdamW displacement of the tensors in $\mathcal{W}_{\mathrm{NPO}}$, including the decoupled weight-decay term, which is therefore orthogonalized jointly with the gradient-based update rather than applied as a separate shrinkage; the other parameters retain their unmodified AdamW updates. All results are the test accuracy at the final epoch; CIFAR results are averaged over three seeds, and ImageNet results use a single seed, following the DeiT protocol. The remaining settings are listed in Appendix C.3.

4.2 FAILURE MECHANISM OF FDFA

FDFA fails to train the non-convolutional models (Section 4.3); here we examine what happens during training. Figure 2 shows the diagnostics for ViT on CIFAR-100 over the first 30 epochs, computed every 250 steps on a fixed diagnostic batch of 32 inputs with reverse-mode AD that is used only for the diagnostics (Appendix C.4).

Skewed updates. We measure the skew of an update by its stable rank $\|D_l\|_F^2/\|D_l\|_2^2$, which equals the rank when the singular values are equal and approaches 1 as the skew grows. FDFA updates have stable rank close to one throughout training, below that of BP (Figure 2a). For IsoFDFA, we measure the norm-rescaled Newton–Schulz updates, which maintain substantially higher stable rank.

Anisotropic and inflated weights. We measure the skew of a weight matrix $W \in \mathbb{R}^{m\times n}$ with singular values $\sigma_1, \ldots, \sigma_q$, $q = \min(m, n)$, by its anisotropy $\mathcal{A}(W) = q\sum_i \sigma_i^4/(\sum_i \sigma_i^2)^2 - 1 = \mathrm{Var}(\sigma^2)/\mathbb{E}[\sigma^2]^2$, which is zero for a full-rank matrix with equal singular values and approaches $q-1$ as the matrix collapses to rank one, and its inflation by the Frobenius norm relative to initialization;

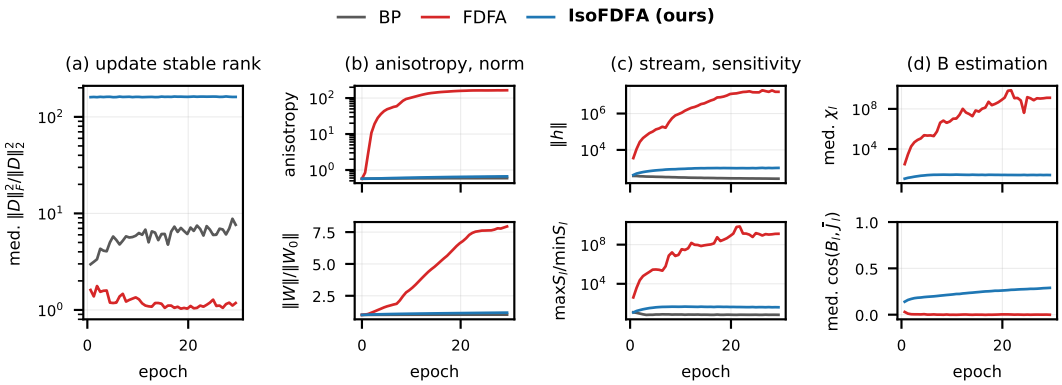

Figure 2: Diagnostics for ViT on CIFAR-100 over the first 30 epochs (BP, FDFA, IsoFDFA): (a) stable rank of the updates, (b) weight anisotropy and norm growth, (c) activation norm and sensitivity imbalance across feedback points, (d) χ_l and alignment of B_l with the mean Jacobian. Definitions are given in the text and the aggregation of each panel in Appendix C.4.

Table 1: Test accuracy (%) on CIFAR-10 and CIFAR-100 (mean $\pm$ std over three seeds). "Passes" is the number of forward-mode passes per step used by the learning rule; it excludes local within-block differentiation and the diagnostic computations of Figure 2. BP + NPO applies our norm-preserving update orthogonalization (NPO) to the Adam step of backpropagation. †Single-pass GrAPE without BP anchor steps. The best backpropagation-free result in each column is in bold.

		CIFAR-10				CIFAR-100			
Method	Passes	ResNet	Mixer	ViT	Swin	ResNet	Mixer	ViT	Swin
BP	–	90.0±1.1	77.3±0.5	70.7±1.4	85.6±0.2	67.2±0.0	42.8±1.2	41.7±0.2	56.4±0.5
BP + NPO	–	90.9±0.4	84.2±0.5	74.1±0.4	87.0±1.2	66.2±0.1	53.9±0.7	44.3±0.0	61.8±0.8
DFA	0	79.6±0.3	23.3±1.6	26.1±1.1	18.5±2.7	39.6±0.5	3.6±0.5	3.9±0.4	2.4±0.2
FG	1	37.0±0.7	27.5±1.3	26.5±0.8	17.8±2.8	9.5±1.8	5.8±1.6	6.9±0.2	1.9±0.5
MTFG ($k = 4$)	4	44.0±1.3	31.7±1.9	28.3±0.8	12.5±2.5	13.6±2.4	7.3±1.6	10.6±0.7	2.1±0.4
FDFA	1	79.5±0.7	19.0±1.8	19.9±5.3	18.2±3.8	44.3±0.5	2.0±0.3	5.0±1.1	2.8±0.5
GrAPE†	1	79.5±0.6	21.8±0.9	22.5±4.3	20.5±1.7	45.8±0.3	2.8±0.5	3.8±0.5	2.5±0.3
IsoFDFA (ours)	1	**83.2**±0.4	**75.9**±0.1	**72.4**±0.4	**66.3**±0.8	**57.5**±0.5	**50.9**±0.4	**45.9**±0.4	**49.7**±0.4

both are averaged over all weight matrices except the classifier. Under FDFA, the anisotropy grows by more than two orders of magnitude within the first 10 epochs, followed by growth of the norm (Figure 2b), matching the order in Section 2.2. Under IsoFDFA, both stay close to those of BP.

Sensitivity imbalance. We measure the inflation of the residual stream, the sum of block outputs carried by the residual connections, by the activation norm $\|h_l\|$ at the feedback points, averaged over the inputs and the points, and the imbalance by the ratio $\max_l S_l / \min_l S_l$ of the sensitivities in Eq. (3). Under FDFA, the activation norm grows by orders of magnitude and the sensitivity ratio reaches 10^9 (Figure 2c); under IsoFDFA, both remain within an order of magnitude of BP.

Feedback estimation. We quantify the relative strength of the contributions of the other points by χ_l in Eq. (3), and separately evaluate the alignment of the learned feedback by the cosine similarity between B_l and the mean Jacobian $\bar{J}_l$ computed with reverse-mode AD on the fixed diagnostic batch. Under FDFA, χ_l explodes together with the sensitivity ratio, and the cosine similarity remains near zero throughout training (Figure 2d). Under IsoFDFA, χ_l stays small and the cosine similarity rises steadily; since $\chi_l = 5$ even for six points of equal sensitivity, this means not that the contributions of the other points vanish but that their explosive growth is prevented.

These results are consistent with the mechanism of Section 2.2. IsoFDFA intervenes only on the skew of the updates, and every later stage is mitigated; in this setting, removing the skew is sufficient to suppress the chain.

4.3 Performance Evaluation

CIFAR. On the non-convolutional models (MLP-Mixer, ViT, and Swin), none of the backpropagation-free baselines we evaluate trains well in our setting (Table 1): on CIFAR-100, the DFA family stays near chance level and the forward-gradient family at most about 11%, with little gain from four passes instead of one. IsoFDFA trains all three models with a single pass; on MLP-Mixer and ViT it is on par with our BP baseline at the same learning rate, while a gap remains on Swin. On these models, BP + NPO also improves over BP, and the operation therefore benefits the optimization itself. The gain of IsoFDFA over FDFA, about 41–49 points, is far larger, however, and points to a problem specific to FDFA. On the convolutional ResNet-18, the existing methods do train, and IsoFDFA still improves over FDFA. CIFAR-10 shows the same picture: IsoFDFA is on par with BP on MLP-Mixer and ViT, while a gap remains on Swin. Figure 1 (bottom) shows BP + NPO as the BP reference, since it applies the same update transformation as IsoFDFA and the two then differ only in how the error is delivered; the learning curves show that FDFA makes no progress from the beginning, whereas IsoFDFA learns steadily and approaches BP.

ImageNet. Under the training configurations in Table 5, IsoFDFA reaches 25.8% top-1 accuracy on DeiT, compared with 0.2% for FDFA (Table 2). A substantial gap to BP remains. To save memory, the feedback matrices on ImageNet are pooled over tokens: the probe at each block output is averaged over its T_l tokens, which shrinks B_l from $C \times T_l d_l$ to $C \times d_l$ for token width d_l, and the delivered error is broadcast to all tokens with a scale of $1/T_l$ (Appendix C.2). For DeiT, this reduces the number of feedback parameters from 0.83 billion to 4.6 million (Appendix C.2). Whether this pooling itself costs accuracy is examined in the next section.

Table 2: Top-1 / top-5 accuracy (%) of DeiT on ImageNet-1k.

Method	Top-1	Top-5
BP	79.9	95.0
FDFA	0.2	0.8
IsoFDFA (ours)	25.8	45.5

4.4 Ablations

We run the ablations of Table 3 on CIFAR-100 with MLP-Mixer and ViT, where the failure of FDFA is pronounced, and discuss groups (i)–(vi) below.

(i) IsoFDFA orthogonalizes the Adam output and restores its Frobenius norm. We compare it with Muon in place of Adam, which orthogonalizes the momentum and applies it at its own fixed scale; with NPO without norm preservation, which orthogonalizes the Adam output but applies it with the Muon scaling rule; and with NPO before Adam, which acts on the gradient instead of the update. Orthogonalizing Adam updates with the Muon scaling rule already improves over FDFA, and restoring the norm of the raw Adam update raises the accuracy substantially further, by 16.6 and 8.1 points. Simply replacing Adam with Muon helps much less, while applying NPO to the gradient comes close to IsoFDFA on MLP-Mixer but falls well short of it on ViT. These results support combining update isotropy with the update magnitude inherited from the optimizer.

(ii) We compare NPO with soft and hard weight corrections applied after Adam, using the same parameter-selection and reshaping rules; both preserve the weight norm immediately after the Adam step (Appendix C.5). Both improve substantially over FDFA, but NPO performs better on both models; controlling the update spectrum is more effective than these weight-level corrections.

(iii) With RMSProp, IsoFDFA also improves substantially over FDFA; the effect is not specific to Adam, although the gain on MLP-Mixer is smaller than with Adam.

(iv) Per-point FDFA improves over single-pass FDFA but remains well below single-pass IsoFDFA: removing cross-point noise alone does not recover the accuracy of IsoFDFA, even with L forward-mode passes per step instead of one.

(v) Pooling changes the accuracy of IsoFDFA only slightly, by 0.9 and 0.1 points, and pooled IsoFDFA still far outperforms pooled FDFA; on CIFAR-100, the memory saving of the pooled feedback used on ImageNet comes at little cost in accuracy.

(vi) One or two Newton–Schulz iterations give little improvement on either model, and accuracy rises sharply between two and five iterations; on MLP-Mixer, five iterations already reach the ac-

Table 3: Ablations on CIFAR-100 (test accuracy %, 100 epochs). *Base* denotes the method to which each modification is applied; rows that alter NPO itself are listed under FDFA. Unless noted, all runs use Adam and 8 Newton–Schulz iterations.

Modification	Base	MLP-Mixer	ViT
–	FDFA	2.0	5.0
–	**IsoFDFA**	**50.9**	**45.9**
(i) Variants of update orthogonalization			
Muon instead of Adam	FDFA	7.2	19.0
NPO without norm preservation (Muon scale)	FDFA	34.3	37.8
NPO applied to the gradient before Adam	FDFA	47.3	31.4
(ii) Weight orthogonalization instead of NPO			
Soft norm-preserving correction	FDFA	28.6	37.7
Hard norm-preserving projection	FDFA	27.0	34.3
(iii) Base optimizer			
RMSProp instead of Adam	FDFA	2.5	2.3
RMSProp instead of Adam	IsoFDFA	42.2	42.7
(iv) Gradient estimation			
Per-point estimation (L passes)	FDFA	23.5	18.3
(v) Feedback matrices			
Pooled feedback matrices	FDFA	3.3	16.2
Pooled feedback matrices	IsoFDFA	50.0	45.8
(vi) Newton–Schulz iterations			
1 / 2 / 5 / 8 iterations	IsoFDFA	3.8 / 6.5 / 52.2 / 50.9	2.5 / 5.2 / 41.9 / 45.9

curacy of eight, whereas ViT still gains from eight. The iteration count therefore matters for the effectiveness of NPO.

5 Limitations and Conclusion

This work showed that controlling update spectra substantially improves single-pass FDFA without changing the feedback-estimation rule. We proposed a failure mechanism in which the spectral concentration of the updates interacts with feedback estimation, and NPO makes the updates isotropic without changing their norm. In the ViT diagnostics, this intervention suppresses weight and activation growth, reduces sensitivity imbalance, and improves feedback alignment. On CIFAR-100, IsoFDFA substantially improves FDFA on MLP-Mixer, ViT, and Swin-T, with accuracy competitive with our BP baselines on MLP-Mixer and ViT. These results use one forward-mode pass per step and no error backpropagation across blocks.

A gap to BP remains, and several factors that NPO does not address may contribute to it. On ImageNet-1k, IsoFDFA raises the top-1 accuracy of DeiT-S from 0.2% under FDFA to 25.8%, still well below the 79.9% of BP, and a gap also remains on Swin-T. NPO changes only the spectrum of the update and does not modify the feedback estimation itself; the probe noise inherent to single-pass estimation and the drift of the estimation target during training therefore persist. On ImageNet, moreover, the feedback matrices are pooled to save memory; pooling costs little accuracy in the CIFAR-100 ablations, but that experiment does not measure its effect on ImageNet. How much each of these three factors contributes to the ImageNet gap has not been quantified and is left for future work. A second question we have not separated is how much of the benefit of NPO comes from resolving a problem specific to FDFA and how much from improving the optimization itself, since NPO also improves BP on the evaluated non-convolutional models. Moreover, our analysis does not establish convergence of the coupled weight and feedback updates, and our experiments are limited to image classification with a common learning rate.

Our results suggest that controlling the update spectrum is an effective complement to learned feedback in single-pass FDFA. Because NPO operates on optimizer-generated updates, it can be combined with improvements to feedback estimation. Combining NPO with lower-variance estimators or improved tracking, and extending the sensitivity diagnostics to other local learning rules, are directions for future work.

AI USE STATEMENT

In this work, we used generative AI tools (large language models) for writing assistance: editing the authors' drafts, revising wording, formatting LaTeX, and preparing bibliography entries. We also used them to write the experiment code, to refine the formulation of the method after it was conceived by the authors, to assist in analyzing the experimental results, to review drafts of the manuscript, and to check derivations. We have not used generative AI tools for research ideation: the research question, the hypothesis, and the proposed method originate from the authors. The authors reviewed all AI-assisted work, including the code, the text, and the bibliography, and adopted AI-suggested refinements and interpretations only after checking them against the derivations and the experimental data. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses only publicly available image-classification benchmarks, collects no new data, and involves no human subjects.

REPRODUCIBILITY STATEMENT

The proposed method is defined in Section 2.3, and the complete procedure for one training step is given in Algorithm 1 (Appendix B). The selection and reshaping rules for the parameters transformed by NPO are stated in Section 2.3 and Appendix B, and the model configurations and the placement of the feedback points, including the pooled variant used on ImageNet, are described in Appendices C.1 and C.2. The main hyperparameters, data augmentations, and numbers of seeds are listed in Section 4.1 and Table 5 in Appendix C.3. The definitions, measurement conditions, and aggregation of the diagnostics in Figure 2 are given in Appendix C.4. The derivations supporting the claims of Sections 2.2 and 2.3 are provided with their assumptions in Appendix A. All datasets used are publicly available, and the only preprocessing is the resizing and augmentation described in Section 4.1. The source code will be released upon acceptance.

REFERENCES

Florian Bacho and Dominique Chu. Low-variance forward gradients using direct feedback alignment and momentum. *Neural Networks*, 169:572–583, January 2024. ISSN 0893-6080. doi: 10.1016/j.neunet.2023.10.051. URL http://dx.doi.org/10.1016/j.neunet.2023.10.051.

Nitin Bansal, Xiaohan Chen, and Zhangyang Wang. Can we gain more from orthogonality regularizations in training deep networks? In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/bf424cb7b0dea050a42b9739eb261a3a-Paper.pdf.

Sergey Bartunov, Adam Santoro, Blake Richards, Luke Marris, Geoffrey E Hinton, and Timothy Lillicrap. Assessing the scalability of biologically-motivated deep learning algorithms and architectures. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/63c3ddcc7b23daa1e42dc41f9a44a873-Paper.pdf.

Atılım Güneş Baydin, Barak A. Pearlmutter, Don Syme, Frank Wood, and Philip Torr. Gradients without backpropagation, 2022. URL https://arxiv.org/abs/2202.08587.

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology, 2024. URL https://arxiv.org/abs/2409.20325.

Gauthier Boeshertz, Razvan Pascanu, and Claudia Clopath. Overcoming rank collapse in feedback alignment, 2026. URL https://arxiv.org/abs/2606.11123.

Paul Caillon, Erwan Fagnou, Blaise Delattre, and Alexandre Allauzen. Scaling direct feedback learning with Jacobian alignment guarantees. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=kasbbmwk3s.

Francis Crick. The recent excitement about neural networks. *Nature*, 337(6203):129–132, January 1989. ISSN 1476-4687. doi: 10.1038/337129a0. URL http://dx.doi.org/10.1038/337129a0.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, June 2009. doi: 10.1109/cvpr.2009.5206848. URL http://dx.doi.org/10.1109/CVPR.2009.5206848.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Ky Fan and A. J. Hoffman. Some metric inequalities in the space of matrices. *Proceedings of the American Mathematical Society*, 6(1):111–116, February 1955. ISSN 0002-9939. doi: 10.1090/s0002-9939-1955-0067841-7. URL http://dx.doi.org/10.1090/S0002-9939-1955-0067841-7.

Katharina Flügel, Daniel Coquelin, Marie Weiel, Charlotte Debus, Achim Streit, and Markus Götz. Beyond backpropagation: Optimization with multi-tangent forward gradients. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, June 2025. doi: 10.1109/ijcnn64981.2025.11227446. URL http://dx.doi.org/10.1109/IJCNN64981.2025.11227446.

Stephen Grossberg. Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science*, 11(1):23–63, January 1987. ISSN 1551-6709. doi: 10.1111/j.1551-6708.1987.tb00862.x. URL http://dx.doi.org/10.1111/j.1551-6708.1987.tb00862.x.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778. IEEE, June 2016. doi: 10.1109/cvpr.2016.90. URL http://dx.doi.org/10.1109/CVPR.2016.90.

N. J. Higham. Matrix nearness problems and applications. *APPLICATIONS OF MATRIX THEORY*, 22, 1989.

Max Jaderberg, Wojciech Marian Czarnecki, Simon Osindero, Oriol Vinyals, Alex Graves, David Silver, and Koray Kavukcuoglu. Decoupled neural interfaces using synthetic gradients. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1627–1635. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/jaderberg17a.html.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL https://kellerjordan.github.io/posts/muon/.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Tim Tsz-Kit Lau, Qi Long, and Weijie Su. PolarGrad: A class of matrix-gradient optimizers from a unifying preconditioning perspective, 2025. URL https://arxiv.org/abs/2505.21799.

Julien Launay, Iacopo Poli, and Florent Krzakala. Principled training of neural networks with direct feedback alignment, 2019. URL https://arxiv.org/abs/1906.04554.

Julien Launay, Iacopo Poli, François Boniface, and Florent Krzakala. Direct feedback alignment scales to modern deep learning tasks and architectures. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 9346–9360. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/69d1fc78dbda242c43ad6590368912d4-Paper.pdf.

Timothy P. Lillicrap, Daniel Cownden, Douglas B. Tweed, and Colin J. Akerman. Random synaptic feedback weights support error backpropagation for deep learning. *Nature Communications*, 7 (1), November 2016. ISSN 2041-1723. doi: 10.1038/ncomms13276. URL http://dx.doi.org/10.1038/ncomms13276.

Timothy P. Lillicrap, Adam Santoro, Luke Marris, Colin J. Akerman, and Geoffrey Hinton. Backpropagation and the brain. *Nature Reviews Neuroscience*, 21(6):335–346, April 2020. ISSN 1471-0048. doi: 10.1038/s41583-020-0277-3. URL http://dx.doi.org/10.1038/s41583-020-0277-3.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9992–10002. IEEE, October 2021. doi: 10.1109/iccv48922.2021.00986. URL http://dx.doi.org/10.1109/ICCV48922.2021.00986.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Arild Nøkland. Direct feedback alignment provides learning in deep neural networks. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/d490d7b4576290fa60eb31b5fc917ad1-Paper.pdf.

Maria Refinetti, Stéphane D'Ascoli, Ruben Ohana, and Sebastian Goldt. Align, then memorise: the dynamics of learning with feedback alignment. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8925–8935. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/refinetti21a.html.

Mengye Ren, Simon Kornblith, Renjie Liao, and Geoffrey Hinton. Scaling forward gradient with local losses. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=JxpBP1JM15-.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, October 1986. ISSN 1476-4687. doi: 10.1038/323533a0. URL http://dx.doi.org/10.1038/323533a0.

Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. MLP-Mixer: An all-MLP architecture for vision. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 24261–24272. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/cba0a4ee5ccd02fda0fe3f9a3e7b89fe-Paper.pdf.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on*

Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, pp. 10347–10357. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/touvron21a.html.

Mark Tuddenham, Adam Prügel-Bennett, and Jonathan Hare. Orthogonalising gradients to speed up neural network optimisation, 2022. URL https://arxiv.org/abs/2202.07052.

Ziao Wang, Kilian Müller, Matthew Filipovich, Julien Launay, Ruben Ohana, Gustave Pariente, Safa Mokaadi, Charles Brossollet, Fabien Moreau, Alessandro Cappelli, Iacopo Poli, Igor Carron, Laurent Daudet, Florent Krzakala, and Sylvain Gigan. Streamlined optical training of large-scale modern deep learning architectures with direct feedback alignment. *Proceedings of the National Academy of Sciences*, 123(20):e2532022123, 2026.

Ross Wightman. PyTorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

Haozhou Zhang. PolarAdamW: Disentangling spectral control and Schur gauge-equivariance in matrix optimisation, 2026. URL https://arxiv.org/abs/2605.07067.

A Derivations for Section 2

Maximal action of a skewed update. For a unit vector v, $\|D_t v\|_2 \le \sigma_1(D_t)$ with equality at the top right singular vector; $\sigma_1(D_t)$ is therefore the maximal action of one step on its input. For a fixed Frobenius norm, $\|D_t\|_F^2 = \sum_{i=1}^{r} \sigma_i(D_t)^2$ implies that $\sigma_1(D_t)$ ranges from $\|D_t\|_F/\sqrt{r}$, attained when all r nonzero singular values are equal, to $\|D_t\|_F$, attained when the update has rank one. Among updates of the same Frobenius norm, the more the energy is concentrated on a few singular directions, the larger the maximal action.

Accumulation along a shared direction. By the triangle inequality, $\|W_T - W_0\|_2 \le \sum_{t=0}^{T-1} \|D_t\|_2$. Equality holds when there is a common pair of unit vectors (u, v) with $u^\top D_t v = \|D_t\|_2$ for all t, that is, when the updates share a top singular pair with the same sign. In that case

$$u^\top (W_T - W_0) v = \sum_{t=0}^{T-1} u^\top D_t v = \sum_{t=0}^{T-1} \|D_t\|_2, \tag{6}$$

and the accumulated change along that direction is the sum of the per-step spectral norms, which grows at least linearly in the number of steps when the per-step spectral norms are bounded away from zero; when dominant directions persist, skewed updates accumulate into anisotropic, enlarged weights.

Probe variance of the single-pass estimate. Fix the model and a mini-batch $\mathcal{B}$ of size N, and let $u_{m,i} \sim \mathcal{N}(0, I_{n_m})$ be independent across points and samples. The estimate $\widehat{J}_l = \frac{1}{N}\sum_i d_i u_{l,i}^\top$ of Eq. (2) splits into the own term $J_{l,i} u_{l,i} u_{l,i}^\top$ and the cross term $X_{l,i} = \sum_{m \ne l} J_{m,i} u_{m,i} u_{l,i}^\top$ for each sample, with $\mathbb{E}_u[\widehat{J}_l] = \frac{1}{N}\sum_i J_{l,i}$. For $u \sim \mathcal{N}(0, I_n)$ and a matrix A, $\mathbb{E}[\|u\|^2 u^\top A u] = (n+2)\,\mathrm{tr}(A)$, hence $\mathbb{E}\|J u u^\top\|_F^2 = (n+2)\|J\|_F^2$ and $\mathbb{E}\|J u u^\top - J\|_F^2 = (n+1)\|J\|_F^2$. The cross term has zero mean, is uncorrelated with the own term, and satisfies $\mathbb{E}\|X_{l,i}\|_F^2 = \mathbb{E}\|u_{l,i}\|^2\, \mathbb{E}\big\|\sum_{m\ne l} J_{m,i} u_{m,i}\big\|^2 = n_l \sum_{m\ne l} \|J_{m,i}\|_F^2$. Independence across samples then gives, with the batch quantities $S_l^{\mathcal{B}} = \frac{1}{N}\sum_i \|J_{l,i}\|_F^2$ and $\chi_l^{\mathcal{B}} = \sum_{m\ne l} S_m^{\mathcal{B}}/S_l^{\mathcal{B}}$,

$$\mathbb{E}_u\big\|\widehat{J}_l - \mathbb{E}_u[\widehat{J}_l]\big\|_F^2 = \frac{1}{N}\Big[(n_l+1)\, S_l^{\mathcal{B}} + n_l \sum_{m\ne l} S_m^{\mathcal{B}}\Big], \tag{7}$$

and dividing by $S_l^{\mathcal{B}}$ yields the ratio $(n_l+1)/N + n_l \chi_l^{\mathcal{B}}/N$ of Eq. (4); the derivation assumes unpooled feedback and no coupling of the forward computation across samples. Per-point estimation has no cross term and retains $(n_l+1)/N$.

Optimality of the norm-preserving update orthogonalization. Let $D_t = U\Sigma V^\top$ be the compact SVD with rank r and $s = \|D_t\|_F$, and let $\mathcal{C}$ be the set of matrices X with $\mathrm{col}(X) = \mathrm{col}(D_t)$, $\mathrm{row}(X) = \mathrm{row}(D_t)$, and $\|X\|_F = s$. Every $X \in \mathcal{C}$ can be written as $X = UMV^\top$ with an invertible $r \times r$ matrix M and $\|M\|_F = s$. Since $\|M\|_2^2 \geq \|M\|_F^2/r = s^2/r$ with equality exactly when all singular values of M equal $s/\sqrt{r}$, the spectral norm on $\mathcal{C}$ is minimized exactly by $M = (s/\sqrt{r})\, P$ with P an $r \times r$ orthogonal matrix. Among these, $\|X - D_t\|_F = \|(s/\sqrt{r})\, P - \Sigma\|_F$ is minimized by the orthogonal polar factor of Σ, that is, $P = I$ (Fan & Hoffman, 1955; Higham, 1989). The unique solution is therefore $\mathcal{T}(D_t) = (s/\sqrt{r})\, UV^\top$.

B Algorithm

Algorithm 1 summarizes one training step of IsoFDFA (Section 2.3).

Algorithm 1 One training step of IsoFDFA

Require: Forward parameters W_t, feedback matrices $B_{l,t}$, optimizer state z_t

1: Draw a mini-batch and independent probes $u_{l,i} \sim \mathcal{N}(0, I_{n_l})$ for $l = 1, \ldots, L$
2: Run one forward-mode pass at W_t to obtain the logits, the output errors g_i, and $d_i = \sum_{m=1}^{L} J_{m,i} u_{m,i}$
3: **for** each feedback point $l = 1, \ldots, L$ **do**
4: $\quad \widehat{J}_l \leftarrow \frac{1}{N} \sum_i d_i u_{l,i}^\top$; update $B_{l,t+1}$ by the feedback optimizer
5: $\quad \delta_{l,i} \leftarrow B_{l,t+1}^\top g_i$; compute G_l within block l by Eq. (1)
6: **end for**
7: Compute the gradient G_{head} of the classification head from g_i
8: $(W_t^{\mathrm{raw}}, z_{t+1}) \leftarrow \mathrm{Opt}(W_t, \{G_l\} \cup \{G_{\mathrm{head}}\}, z_t)$; $\quad W_{t+1} \leftarrow W_t^{\mathrm{raw}}$
9: **for** each $W \in \mathcal{W}_{\mathrm{NPO}}$ with $D_t = W_t^{\mathrm{raw}} - W_t \neq 0$ **do**
10: $\quad Q_t \leftarrow$ Newton–Schulz polar factor of D_t (Muon coefficients, eight iterations)
11: $\quad W_{t+1} \leftarrow W_t + \|D_t\|_F\, Q_t / \|Q_t\|_F$
12: **end for**
13: **return** W_{t+1}, $B_{l,t+1}$, z_{t+1}

All local gradients in Algorithm 1 correspond to the same pre-update forward state, and the feedback update is performed before delivery. For the pooled feedback used on ImageNet, lines 4 and 5 use the token-averaged probe $\bar{u}_{l,i}$ in place of $u_{l,i}$, and the delivered error $B_{l,t+1}^\top g_i / T_l$ is copied to every token (Appendix C.2). The unbiasedness of Eq. (2) concerns the mean Jacobian estimate, not the parameter gradient in Eq. (1); even an exact mean Jacobian generally differs from the sample-specific Jacobians required by BP.

Parameters transformed by NPO. The set $\mathcal{W}_{\mathrm{NPO}}$ of Section 2.3 consists of the forward-model parameter tensors with two or four dimensions, excluding the classification head and position-dependent bias tables such as the relative-position bias of Swin. Two-dimensional weights are transformed as they are. Four-dimensional convolution kernels of shape $(\mathrm{out}, \mathrm{in}, k, k)$ are reshaped to $(\mathrm{out}, \mathrm{in} \cdot k \cdot k)$ matrices before the transformation and reshaped back afterward. One-dimensional parameters (biases and normalization scales and shifts), the class token, the positional embeddings, and the feedback matrices B_l are not transformed and receive their ordinary optimizer updates. The resulting set contains 25 tensors for the ViT, 33 for MLP-Mixer, and 52 for Swin-T; the same selection and reshaping rules are used for Muon and for the soft and hard weight-orthogonalization controls.

C Experimental Details

C.1 Models

Table 4 lists the configurations of the models used in the experiments; all take 224×224 inputs. ResNet-18 is the standard configuration (He et al., 2016). The ViT is the small configuration of Flügel et al. (2025), not ViT-S/16. Swin-T is the standard configuration (Liu et al., 2021). DeiT-S/16

has the same architecture as ViT-S/16 and is trained with drop-path rate 0.1 and the initialization of `timm` (Wightman, 2019), following Touvron et al. (2021).

Table 4: Model configurations. Depth is the number of blocks (per stage for Swin-T), and MLP is the hidden width of the MLP (ratio to the width in parentheses; token / channel MLP for MLP-Mixer).

Model	Width	Depth	Heads	MLP	Patch
ResNet-18	64–512	4 stages × 2	–	–	–
MLP-Mixer (Mixer-S/16)	512	8	–	256 / 2048	16
ViT (Flügel et al., 2025)	256	6	4	512 (×2)	16
Swin-T	96	(2, 2, 6, 2)	(3, 6, 12, 24)	×4, window 7	4
DeiT-S/16	384	12	6	1536 (×4)	16

ResNet-18. We use the standard ImageNet architecture (He et al., 2016) (7×7 stride-2 stem convolution followed by max-pooling, four stages of two BasicBlocks with 64–512 channels, global average pooling, and a linear classifier) with BatchNorm, applied to CIFAR images resized to 224×224; we do not use the CIFAR-specific 3×3 stem without pooling.

MLP-Mixer. We use the Mixer-S/16 configuration (Tolstikhin et al., 2021): patch size 16 on 224×224 inputs (196 tokens), hidden width 512, 8 mixer blocks, token-mixing MLP width 256 and channel-mixing MLP width 2048 (GELU), followed by a LayerNorm, global average pooling over tokens, and a linear classifier.

C.2 Feedback Points

Feedback points. Every model is split into blocks, and one feedback matrix $B_l \in \mathbb{R}^{C \times n_l}$ is attached to the output of each block except the classifier, where n_l is the dimension of the flattened block output; the classifier receives the true error. For the ViT (6 blocks), MLP-Mixer (8 blocks), and Swin-T (12 blocks over four stages) a block is one transformer or mixer block, and a feedback point sits at every block output (the token map, $n_l = T_l d_l$ for T_l tokens of width d_l); the last block also includes the final LayerNorm and pooling, and its feedback point is the pooled feature ($n_l = d_l$). For ResNet-18 a block is one BasicBlock, giving 8 feedback points: two each on feature maps of size $64 \times 56 \times 56$, $128 \times 28 \times 28$, and $256 \times 14 \times 14$, one on $512 \times 7 \times 7$, and the last on the globally pooled 512-dimensional feature after the final BasicBlock. The first block additionally contains the patch or stem embedding.

Pooled feedback (ImageNet). For DeiT on ImageNet, the feedback matrices at the token points are pooled over tokens. Without pooling, the block output of shape $[T_l, d_l]$ is flattened, B_l is $C \times T_l d_l$ ($1000 \times 75{,}648$ for DeiT-S), the estimate is $\widehat{J}_l = \frac{1}{N} \sum_i d_i u_{l,i}^\top$ of Eq. (2) with the flattened probe $u_{l,i}$, and the delivered error $B_l^\top g_i$ differs across tokens. With pooling, the probe is averaged over the T_l tokens to $\bar{u}_{l,i} \in \mathbb{R}^{d_l}$, B_l shrinks to $C \times d_l$ (1000×384), and the estimate becomes $\widehat{J}_l = \frac{1}{N} \sum_i d_i \bar{u}_{l,i}^\top$. With independent probes across tokens, the expectation of this outer-product estimator equals the token-averaged Jacobian $\bar{J}_l^{\mathrm{tok}} = \mathbb{E}_i[\frac{1}{T_l} \sum_{q=1}^{T_l} J_{l,q,i}]$, which is therefore the target of B_l. The delivered error is a single d_l-dimensional vector broadcast to all tokens, $\delta_{l,i,q} = \alpha_l B_l^\top g_i$, with $\alpha_l = 1/T_l$ in all reported pooled-feedback experiments. This factor is a delivery-scale convention, not a factor required for unbiased estimation. For a constant rescaling of the gradient history, Adam cancels it when $\epsilon = 0$; with finite ϵ, the rescaling changes its effective value from ϵ to $T_l \epsilon$. The last block, whose feedback point is the pooled feature, is unchanged. The total number of feedback parameters drops from $11 \times 75.6\mathrm{M} + 0.38\mathrm{M} \approx 832\mathrm{M}$ (3.3 GB in fp32, about 38 times the 22M parameters of DeiT-S) to $12 \times 0.38\mathrm{M} = 4.6\mathrm{M}$ (18 MB, about 0.2 times the model).

C.3 Training Settings

Table 5 lists the training settings. On CIFAR, the same constant learning rate is used for every method and model; no per-method tuning is performed. On ImageNet, we follow the official DeiT recipe (Touvron et al., 2021).

Table 5: Training settings. The ImageNet augmentation and regularization follow the DeiT recipe (Touvron et al., 2021).

	CIFAR-10 / CIFAR-100	ImageNet-1k
Forward-weight optimizer	Adam ($\beta = 0.9, 0.999$)	AdamW
Learning rate	10^{-4} (all methods and models)	BP: 10^{-3} ($5 \times 10^{-4} \times$ batch/512); FDFA and IsoFDFA: 2.5×10^{-4}
Weight decay	0	0.05 (0 for 1-D parameters, `pos_embed`, `cls_token`; LayerNorm included for FDFA and IsoFDFA)
Schedule	constant	cosine, 5 warm-up epochs
Feedback optimizer	Adam, 10^{-4}; B_l initialized to 0	AdamW with 5 warm-up epochs; B_l initialized to 0
Feedback points	every block output	every block output, pooled
Epochs / batch size	100 / 128	300 / 1024 (8 GPUs $\times$ 128)
Input	224×224 (CIFAR upsampled)	224×224
Augmentation	random resized crop, horizontal flip	Mixup/CutMix, RandAugment, repeated augmentation
Regularization	–	label smoothing, drop-path 0.1
Newton–Schulz	8 iterations, fp32, Muon coefficients	8 iterations, fp32, Muon coefficients
Evaluation	final epoch	final epoch, non-EMA weights
Seeds	3	1

Newton–Schulz iteration and ablation optimizers. The Newton–Schulz iteration uses the coefficients $(3.4445, -4.7750, 2.0315)$ of Muon, normalizes its input by the Frobenius norm, transposes it to be wide, and runs in fp32; the same routine is used for NPO, the hard weight projection, and Muon. In the ablations, Muon is applied to the matrices in $\mathcal{W}_{\mathrm{NPO}}$ and Adam to the other parameters, with learning rate 0.002, Nesterov-style momentum 0.95, scale $\sqrt{\max(1, m/n)}$, and five Newton–Schulz iterations. RMSProp uses learning rate 10^{-4}, $\alpha = 0.99$, $\epsilon = 10^{-8}$, and no momentum.

C.4 Diagnostics

Every 250 training steps, we compute per-sample Jacobians of the logits with respect to the activations at the feedback points on a fixed set of $M = 32$ diagnostic inputs, selected as the first 32 samples of a batch obtained at initialization; these inputs are not excluded from training. All $C = 100$ output coordinates are differentiated with vectorized reverse-mode AD. This reverse-mode AD is used only for the diagnostics, never for the feedback estimation or the parameter updates. Diagnostics are run with the model in training mode, BatchNorm statistics frozen, and dropout disabled. The reported sensitivities and mean Jacobians are the empirical averages $\widehat{S}_l = \frac{1}{M}\sum_{i=1}^{M} \|J_{l,i}\|_F^2$ and $\widehat{\bar{J}}_l = \frac{1}{M}\sum_{i=1}^{M} J_{l,i}$ over these inputs, not the population quantities S_l and $\bar{J}_l$; M is distinct from the training batch size $N = 128$. The panels of Figure 2 aggregate as follows: (a) the median over the update matrices in $\mathcal{W}_{\mathrm{NPO}}$; (b) the simple mean over all matrices in $\mathcal{W}_{\mathrm{NPO}}$ of the anisotropy and of the per-matrix norm ratio $\|W_t\|_F / \|W_0\|_F$; (c, top) the activation norm averaged over the diagnostic inputs at each feedback point and then over the six points; (c, bottom) $\max_l \widehat{S}_l / \min_l \widehat{S}_l$ over the six points; (d) each quantity computed per point and the median over the six points.

C.5 Weight-Orthogonalization Controls

The two controls of Section 4.4 (ii) orthogonalize the weights instead of the update. The soft version applies a norm-preserving weight correction after each Adam step rather than adding a penalty to the loss: with $A_t = W_t^{\mathrm{raw}} \in \mathbb{R}^{m \times n}$ and $c_t^2 = \|A_t\|_F^2 / q$, $q = \min(m, n)$, it computes

$$\widetilde{A}_t = A_t - \eta_{\mathrm{ortho}}\big(A_t A_t^\top A_t - c_t^2 A_t\big), \qquad W_{t+1} = \|A_t\|_F \, \frac{\widetilde{A}_t}{\|\widetilde{A}_t\|_F}, \tag{8}$$

with $\eta_{\mathrm{ortho}} = 0.1$; the norm preserved is that of the weights immediately after the Adam step, not the initial norm or the norm of the update. The hard version sets $W_{t+1} = \|A_t\|_F \, Q_t / \|Q_t\|_F$ after every step, where Q_t is the Newton–Schulz approximation of the polar factor of A_t; normalizing

by $\|Q_t\|_F$ rather than $\sqrt{q}$ keeps the norm exact under the approximate iteration. Both use the same parameter set and reshaping as NPO.