ON THE ROLE OF BALANCED INITIALIZATION IN FEEDBACK ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Feedback alignment (FA) and its variants replace exact backward weights with fixed random feedback matrices, and have recently been argued to retain important properties of gradient-based learning. We identify a fundamental distinction between these dynamics through balancedness, a key invariant of gradient flow in deep linear networks. We show that Random Feedback Alignment (RFA) generically fails to preserve balancedness: even from exactly balanced initialization, the balancedness defect departs from zero. We prove, under mild nondegeneracy conditions, that exact balancedness can occur only at isolated times. Surprisingly, balanced structure nevertheless remains beneficial: in nonlinear networks, balanced initialization improves RFA training over Xavier initialization, and this advantage persists after matching the layerwise norms of the Xavier initialization. These results separate the benefits of balanced structure from its exact conservation, and expose a structural difference between feedback alignment and backpropagation.

1 INTRODUCTION

Artificial neural networks draw their inspiration from networks of biological neurons and have long been a workhorse of modern deep learning (Rumelhart et al., 1986b; Medler, 1998; Elman et al., 1996). Backpropagation (BP) (Rumelhart et al., 1986a) underpins this success. It suffers, however, from three well-known limitations.

First, BP cannot be parallelized: the gradient at layer l depends on the gradient at layer $l+1$, so the backward pass is strictly sequential (Jaderberg et al., 2017). Second, BP is not biologically plausible (Crick, 1989; Marblestone et al., 2016). Its most serious objection is the weight transport problem: BP relies on backward symmetric weight updates, requiring the feedback pathway to use the transpose of the forward weights, for which no biological mechanism is known (Grossberg, 1987). Third, BP requires a full forward pass before parameters can be updated during the backward pass, a phenomenon known as update locking, which imposes a significant memory overhead (Czarnecki et al., 2017; Jaderberg et al., 2017).

A number of biologically motivated alternatives have been proposed. Feedback Alignment (FA) (Lillicrap et al., 2016) replaces the transposed forward weights in the backward pass with a fixed random matrix, and shows that learning still occurs. Direct Feedback Alignment (DFA) (Nøkland, 2016) propagates the output error directly to every hidden layer, addressing the parallelization issue. Direct Random Target Projection (DRTP) (Frenkel et al., 2019) replaces the output error with the target vector, addressing update locking. Sign-concordant feedback (Liao et al., 2016; Xiao et al., 2018; Moskovitz et al., 2018) uses the signs of the forward weights in the feedback path and performs on par with BP. Other methods start with random feedback and train the feedback along with the forward weights, achieving performance closer to BP (Akrout et al., 2019; Kunin et al., 2020; Hanut & Kadmon, 2025).

Empirically, these methods succeed on a range of tasks: DFA can scale to modern architectures including transformers (Launay et al., 2020). They also fail (Liao et al., 2016; Xiao et al., 2018; Bartunov et al., 2018; Launay et al., 2019). The failure on deep convolutional networks has been attributed to a bottleneck effect: convolutional layers lack the degrees of freedom necessary to align with random feedback matrices (Refinetti et al., 2021; Launay et al., 2019).

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.

Why do these methods work at all? The standard answer is alignment: forward weights gradually align with the feedback matrices, which produces gradient alignment, and this drives learning (Lillicrap et al., 2016). Refinetti et al. (2021) showed that learning with DFA proceeds in two stages, an alignment phase followed by a memorization phase. Subsequent work has complicated this picture. Chu & Bacho (2024) argued that gradient alignment is necessary but not sufficient, and modeled alignment as a byproduct of a random walk under RFA. Song et al. (2021) showed that alignment does not occur in the over-parameterized regime, but that regularization restores it. Boeshertz et al. (2026) observed that when the weights start to align, the effective rank of the gradients decreases markedly, confining updates to a low-dimensional subspace that can prevent alignment from continuing. Caillon et al. (2026) proposed GrAPE, which computes a local cosine-alignment loss with forward-gradient estimates to realign each layer's feedback matrix toward Jacobian-aligned directions. Lee & Lee (2026) showed that DFA fails in fine-tuning, analyzed the failure through weight and gradient alignment, and recovered performance by reconstructing the feedback matrix from the pre-trained weights.

In parallel to this alignment literature, a separate line of work studies invariants and convergence. Robertson & Koyejo (2023) derived a conservation law analogous to the norm-balancing invariant of Du et al. (2018) for feedback alignment, coupling the forward weights to the feedback matrix, and used this law to establish sufficient conditions for layer-wise alignment. Robertson & Koyejo (2024) extended this analysis and introduced an alignment dominance condition under which convergence guarantees can be derived. Their guarantee for sign-concordant feedback requires a norm condition on the initialization and applies to two-layer networks with leaky ReLU activations, exponential loss, and data that is orthogonally separable or nearly orthogonal. Their guarantee for standard feedback alignment requires, in addition, an initialization in which the feedback weights are set equal to the output weights at the start of training. They argue that FA shares the implicit regularization of gradient descent, and therefore is not as structurally different from it as previously assumed. However, their results do not cover generic Random Feedback Alignment, where the feedback matrix is drawn at random and has no relation to the forward weights.

In this paper, we investigate an invariant that operates independently of the feedback matrix. Arora et al. (2019) previously identified balancedness—the matching of Gram matrices across adjacent layers—as an invariant of gradient descent, demonstrating that it ensures linear convergence in linear networks and empirically accelerates training in non-linear architectures. Because balancedness relies exclusively on forward weights, it bypasses the need for coordination between forward and feedback pathways, rendering it highly attractive from a biological perspective. However, we prove that under generic Random Feedback Alignment, balancedness fundamentally fails to hold as an invariant.

Surprisingly, despite this theoretical breakdown, we empirically observe that the benefits of balanced initialization carry over to RFA, much like in backpropagation. On CIFAR-10 and MNIST, balanced initialization performs better than Xavier.

2 Setup

In this section, we describe the continuous time dynamics of a two-layer linear network trained with Random Feedback Alignment. Then, we formally define the notion of δ-balancedness.

2.1 Discrete RFA Updates and Continuous Dynamics

We consider a two-layer linear network mapping an input vector $\mathbf{x} \in \mathbb{R}^n$ to an output vector $\mathbf{y} \in \mathbb{R}^n$ via hidden activations $\mathbf{h} \in \mathbb{R}^m$:

$$\mathbf{h} = \mathbf{W}_1\mathbf{x}, \qquad \mathbf{y} = \mathbf{W}_2\mathbf{h}, \tag{1}$$

where $\mathbf{W}_1 \in \mathbb{R}^{m\times n}$ and $\mathbf{W}_2 \in \mathbb{R}^{n\times m}$ represent trainable weight matrices. The target behavior is governed by a fixed linear operator $\mathbf{T} \in \mathbb{R}^{n\times n}$ such that $\mathbf{y}^* = \mathbf{T}\mathbf{x}$. We define the residual operator matrix as $\mathbf{E} := \mathbf{T} - \mathbf{W}_2\mathbf{W}_1$, yielding the prediction error vector $\mathbf{e} = \mathbf{E}\mathbf{x}$.

Under standard backpropagation, parameter updates follow gradient descent on the squared error loss $\mathcal{L} = \frac{1}{2}\|\mathbf{y} - \mathbf{y}^*\|_2^2$. For a single input sample $\mathbf{x}$ and step size $\eta > 0$, the updates are given by

$$\Delta\mathbf{W}_2 = \eta\mathbf{E}\mathbf{x}\mathbf{x}^\top\mathbf{W}_1^\top, \qquad \Delta\mathbf{W}_1 = \eta\mathbf{W}_2^\top\mathbf{E}\mathbf{x}\mathbf{x}^\top. \tag{2}$$

To resolve the weight transport problem inherent in using $\mathbf{W}_2^\top$, Random Feedback Alignment (Lillicrap et al., 2016) replaces the exact transpose in the first-layer update with a fixed, random matrix $\mathbf{B} \in \mathbb{R}^{m \times n}$:

$$\Delta \mathbf{W}_2 = \eta \mathbf{E}\mathbf{x}\mathbf{x}^\top \mathbf{W}_1^\top, \qquad \Delta \mathbf{W}_1 = \eta \mathbf{B}\mathbf{E}\mathbf{x}\mathbf{x}^\top. \tag{3}$$

Under the whitened input assumption ($\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \mathbf{I}_n$), the expected parameter updates simplify to the below equations(Lillicrap et al., 2016):

$$\Delta \mathbf{W}_2 = \eta \mathbf{E}\mathbf{W}_1^\top, \qquad \Delta \mathbf{W}_1 = \eta \mathbf{B}\mathbf{E}. \tag{4}$$

Letting $\mathbf{W}_1^{(k)}$ and $\mathbf{W}_2^{(k)}$ denote the weight matrices at iteration step $k \in \mathbb{N}$, equation equation 4 defines a discrete Forward Euler update with step size η:

$$\frac{\mathbf{W}_2^{(k+1)} - \mathbf{W}_2^{(k)}}{\eta} = \mathbf{E}^{(k)} \left(\mathbf{W}_1^{(k)}\right)^\top, \qquad \frac{\mathbf{W}_1^{(k+1)} - \mathbf{W}_1^{(k)}}{\eta} = \mathbf{B}\mathbf{E}^{(k)}. \tag{5}$$

As rigorously established in B, as the step size vanishes ($\eta \to 0$), these discrete Euler iterates converge uniformly over finite time intervals to the continuous-time matrix differential equations:

$$\dot{\mathbf{W}}_2(t) = \mathbf{E}(t)\mathbf{W}_1(t)^\top, \qquad \dot{\mathbf{W}}_1(t) = \mathbf{B}\mathbf{E}(t). \tag{6}$$

To analyze system equation 6, we stack the flattened weight matrices into a single joint state vector $\boldsymbol{\phi}(t) := [\mathrm{vec}(\mathbf{W}_1(t))^\top, \mathrm{vec}(\mathbf{W}_2(t))^\top]^\top \in \mathbb{R}^{2mn}$. The system reduces to the autonomous vector field:

$$\dot{\boldsymbol{\phi}}(t) = \mathbf{f}(\boldsymbol{\phi}(t)), \quad \boldsymbol{\phi}(0) = \boldsymbol{\phi}_0, \tag{7}$$

where $\mathbf{f} \colon \mathbb{R}^{2mn} \to \mathbb{R}^{2mn}$ is a polynomial map defined by $\mathbf{f}(\boldsymbol{\phi}) := [\mathrm{vec}(\mathbf{B}\mathbf{E})^\top, \mathrm{vec}(\mathbf{E}\mathbf{W}_1^\top)^\top]^\top$

2.2 δ-BALANCEDNESS

Definition 1 (δ-Balancedness)**.** A two-layer weight sequence $(\mathbf{W}_1, \mathbf{W}_2)$ is said to be δ-*balanced* for some parameter $\delta \geq 0$ if

$$\|\mathbf{W}_2^\top \mathbf{W}_2 - \mathbf{W}_1 \mathbf{W}_1^\top\|_F \leq \delta, \tag{8}$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

Arora et al. (2019) introduced a balanced initialization algorithm for deep linear networks. They propose an algorithm based on SVD to make the weights balanced at initialization. Under gradient descent, fully-connected linear networks whose weights are δ-balanced at initialization remain δ-balanced throughout training. In particular, if the weights are 0-balanced at initialization, they remain 0-balanced throughout. This property was used to prove linear convergence in linear networks. It also serves as a necessary condition, as adversarial unbalanced initializations shown by Arora et al. (2019) can make the gradient descent struggle. Furthermore, empirical observations showed that the convergence benefits of such an initialization carry over to non-linear settings as well. In the next section we formally prove that RFA training dynamics fail to preserve this invariant

3 GRADIENT FLOW DYNAMICS AND ANALYTICITY

In this section, we prove our main theorem. We quantify the *inter-layer balance* at time t by

$$\mathbf{D}(t) := \mathbf{W}_2^\top(t)\mathbf{W}_2(t) - \mathbf{W}_1(t)\mathbf{W}_1^\top(t), \tag{9}$$

We invoke the Cauchy–Kovalevski theorem (see Appendix A) to establish that the unique solution to equation 7 is real-analytic, which implies that $\mathbf{D}(t)$ is also real-analytic in t. Next, using a Taylor–Peano expansion, we analyze the local behavior of $\mathbf{D}(t)$ for small t. Combining this local expansion with the real-analyticity of $\mathbf{D}(t)$, we invoke the Identity Theorem (see Appendix A) to conclude that $\mathbf{D}(t) = 0$ can occur only at isolated time points.

Lemma 1 (Global Real-Analyticity of $\mathbf{D}(t)$)**.** *Let $\boldsymbol{\phi}(t) := [\mathrm{vec}(\mathbf{W}_1(t))^\top, \mathrm{vec}(\mathbf{W}_2(t))^\top]^\top$ be the unique classical C^1 solution to system equation 7 on its maximal interval of existence $[0, T_{\max})$. Then, the solution trajectory $\boldsymbol{\phi}(t)$ is real-analytic on $[0, T_{\max})$. Consequently, the weight matrices $\mathbf{W}_1(t)$ and $\mathbf{W}_2(t)$ and the inter-layer balance $\mathbf{D}(t)$ are globally real-analytic on $[0, T_{\max})$.*

Proof. The vector field $\mathbf{f}$ governing the continuous-time RFA dynamics equation 7 is a polynomial in the weight matrices, and is therefore real-analytic on $\mathbb{R}^{2mn}$. This property allows us to leverage the Cauchy–Kovalevski theorem to establish strong regularity guarantees for the network's training trajectory.

To establish the real-analyticity of ϕ, we first apply Cauchy–Kovalevski (Appendix A) at the origin. This yields a locally real-analytic solution on some interval $(-\epsilon_0, \epsilon_0)$. By the uniqueness of solutions (Appendix B), ϕ coincides with this local solution on $[0, \epsilon_0)$, so ϕ is real-analytic on this initial interval.

Define

$$T^* := \sup\{t \in [0, T_{\max}) : \phi \text{ is real-analytic on } [0, t]\}.$$

Clearly $T^* \geq \epsilon_0 > 0$. Suppose for contradiction that $T^* < T_{\max}$. Since $T^* < T_{\max}$, the solution ϕ exists and is C^1 at T^*. Applying Cauchy–Kovalevski at $t = T^*$ yields a real-analytic solution ψ on some neighborhood $(T^* - \epsilon, T^* + \epsilon)$ with $\epsilon > 0$. By uniqueness(applied in backward time), ψ coincides with ϕ on the overlap $(T^* - \epsilon, T^*)$, and hence extends ϕ analytically to $[0, T^* + \epsilon)$. This contradicts the definition of T^*. Therefore $T^* = T_{\max}$, proving that ϕ is real-analytic on the entire maximal interval $[0, T_{\max})$.

As a direct consequence, all entries of $\mathbf{W}_1(t)$ and $\mathbf{W}_2(t)$ are globally real-analytic on $[0, T_{\max})$. Since matrix transposition, multiplication, and subtraction consist of finite sums and products of these entries, each entry of $\mathbf{D}(t)$ is a polynomial composition of real-analytic functions, and is therefore globally real-analytic on $[0, T_{\max})$. Furthermore, the discrete FA iterates are guaranteed to converge uniformly to ϕ on any compact subinterval $[0, T] \subset [0, T_{\max})$ (see Appendix B). □

Lemma 2 (Local Behavior of $\mathbf{D}(t)$ at Initialization)**.** *Consider the continuous RFA system equation 6 under zero initialization $\mathbf{W}_1(0) = \mathbf{0}$ and $\mathbf{W}_2(0) = \mathbf{0}$. For small $t > 0$, the discrepancy matrix $\mathbf{D}(t) := \mathbf{W}_2(t)^\top \mathbf{W}_2(t) - \mathbf{W}_1(t)\mathbf{W}_1(t)^\top$ satisfies the Taylor expansion:*

$$\mathbf{D}(t) = -\mathbf{B}\mathbf{T}\mathbf{T}^\top\mathbf{B}^\top t^2 + o(t^2). \tag{10}$$

In particular, if $\mathbf{BT} \neq \mathbf{0}$, then $\mathbf{D}(t) \neq \mathbf{0}$ for sufficiently small $t > 0$.

Proof. Under zero initialization ($\mathbf{W}_1(0) = \mathbf{0}$ and $\mathbf{W}_2(0) = \mathbf{0}$), the initial residual matrix reduces to $\mathbf{E}(0) = \mathbf{T} - \mathbf{W}_2(0)\mathbf{W}_1(0) = \mathbf{T}$. Evaluating the system derivatives at $t = 0$ via the flow equations equation 6 gives:

$$\dot{\mathbf{W}}_1(0) = \mathbf{B}\mathbf{E}(0) = \mathbf{B}\mathbf{T}, \tag{11}$$

$$\dot{\mathbf{W}}_2(0) = \mathbf{E}(0)\mathbf{W}_1(0)^\top = \mathbf{0}. \tag{12}$$

Differentiating the residual matrix $\mathbf{E}(t) = \mathbf{T} - \mathbf{W}_2(t)\mathbf{W}_1(t)$ with respect to t yields $\dot{\mathbf{E}}(t) = -\dot{\mathbf{W}}_2(t)\mathbf{W}_1(t) - \mathbf{W}_2(t)\dot{\mathbf{W}}_1(t)$, which evaluates at $t = 0$ to $\dot{\mathbf{E}}(0) = \mathbf{0}$. Taking second-order derivatives of the weight trajectories yields:

$$\ddot{\mathbf{W}}_1(0) = \mathbf{B}\dot{\mathbf{E}}(0) = \mathbf{0}, \tag{13}$$

$$\ddot{\mathbf{W}}_2(0) = \dot{\mathbf{E}}(0)\mathbf{W}_1(0)^\top + \mathbf{E}(0)\dot{\mathbf{W}}_1(0)^\top = \mathbf{T}(\mathbf{B}\mathbf{T})^\top = \mathbf{T}\mathbf{T}^\top\mathbf{B}^\top. \tag{14}$$

Applying Taylor's theorem with Peano remainder around $t = 0$, the weight trajectories expand as:

$$\mathbf{W}_1(t) = \mathbf{B}\mathbf{T}t + o(t^2), \tag{15}$$

$$\mathbf{W}_2(t) = \frac{1}{2}\mathbf{T}\mathbf{T}^\top\mathbf{B}^\top t^2 + o(t^2). \tag{16}$$

Substituting these expansions into $\mathbf{D}(t) = \mathbf{W}_2(t)^\top\mathbf{W}_2(t) - \mathbf{W}_1(t)\mathbf{W}_1(t)^\top$, we obtain:

$$\mathbf{W}_1(t)\mathbf{W}_1(t)^\top = \big(\mathbf{B}\mathbf{T}t + o(t^2)\big)\big(\mathbf{B}\mathbf{T}t + o(t^2)\big)^\top = \mathbf{B}\mathbf{T}\mathbf{T}^\top\mathbf{B}^\top t^2 + o(t^2), \tag{17}$$

$$\mathbf{W}_2(t)^\top\mathbf{W}_2(t) = \left(\frac{1}{2}\mathbf{T}\mathbf{T}^\top\mathbf{B}^\top t^2 + o(t^2)\right)^\top \left(\frac{1}{2}\mathbf{T}\mathbf{T}^\top\mathbf{B}^\top t^2 + o(t^2)\right) = o(t^2). \tag{18}$$

Subtracting the two terms yields:

$$\mathbf{D}(t) = -\mathbf{B}\mathbf{T}\mathbf{T}^\top\mathbf{B}^\top t^2 + o(t^2). \tag{19}$$

The matrix $\mathbf{B}\mathbf{T}\mathbf{T}^\top\mathbf{B}^\top = (\mathbf{B}\mathbf{T})(\mathbf{B}\mathbf{T})^\top$ is positive semidefinite and is strictly non-zero if and only if $\mathbf{BT} \neq \mathbf{0}$. For a generic random matrix $\mathbf{B}$ and non-zero target $\mathbf{T}$, the condition $\mathbf{BT} \neq \mathbf{0}$ holds almost surely. Thus $\mathbf{D}(t) \neq \mathbf{0}$ for sufficiently small $t > 0$. □

Table 1: Layer-wise evaluation of structural coupling metrics across network depth.

	Balanced Initialization		Xavier Initialization	
Index / Layer	**Init**	**Train**	**Init**	**Train**
1	1.0000 ± 0.0000	0.0348 ± 0.0436	0.0254 ± 0.0193	0.0256 ± 0.0190
2	1.0000 ± 0.0000	0.0789 ± 0.1225	0.0383 ± 0.0276	0.0353 ± 0.0261
3	1.0000 ± 0.0000	0.1306 ± 0.1553	0.0559 ± 0.0394	0.0529 ± 0.0444
4	1.0000 ± 0.0000	0.1053 ± 0.0409	0.0751 ± 0.0478	0.0850 ± 0.0417

Theorem 1 (Absence of Balancedness under RFA)**.** *Let* $(\mathbf{W}_1(t), \mathbf{W}_2(t))$ *evolve according to the continuous RFA dynamics equation 6 under zero initialization on the maximal interval of existence* $[0, T_{\max})$*. Assume* $\mathbf{BT} \neq \mathbf{0}$*. Then, exact layer balance:*

$$\mathbf{W}_2(t)^\top \mathbf{W}_2(t) = \mathbf{W}_1(t)\mathbf{W}_1(t)^\top \quad (\textit{i.e.,}\ \mathbf{D}(t) = \mathbf{0}) \tag{20}$$

is not a conserved quantity of the system. Furthermore, exact balance holds at most on a discrete set of isolated time points in $[0, T_{\max})$ *and cannot persist over any interval of positive length.*

Proof. By Lemma 1, the solution trajectory is globally real-analytic on $I = [0, T_{\max})$. Consequently, every entry $D_{ij}(t)$ of the discrepancy matrix $\mathbf{D}(t) := \mathbf{W}_2(t)^\top \mathbf{W}_2(t) - \mathbf{W}_1(t)\mathbf{W}_1(t)^\top$ is a real-analytic function of t on $[0, T_{\max})$.

Next, by Lemma 2, the local Taylor expansion of $\mathbf{D}(t)$ near $t = 0$ is given by:

$$\mathbf{D}(t) = -\mathbf{BTT}^\top \mathbf{B}^\top t^2 + o(t^2). \tag{21}$$

Because $\mathbf{BT} \neq \mathbf{0}$, the leading coefficient matrix $-\mathbf{BTT}^\top \mathbf{B}^\top = -(\mathbf{BT})(\mathbf{BT})^\top$ is non-zero. Thus, there exists at least one entry (i, j) for which $D_{ij}(t) \neq 0$ for all $t \in (0, \epsilon)$ for some sufficiently small $\epsilon > 0$. This establishes that $\mathbf{D}(t)$ is not identically zero on any open right-neighborhood of $t = 0$.

Now, define the zero set of the discrepancy matrix on the domain of existence:

$$\mathcal{Z} := \{t \in [0, T_{\max}) : \mathbf{D}(t) = \mathbf{0}\}. \tag{22}$$

If $\mathcal{Z}$ contained an accumulation point in $[0, T_{\max})$, the Identity Theorem for real-analytic functions (Appendix A) would imply that $\mathbf{D}(t) \equiv \mathbf{0}$ identically for all $t \in [0, T_{\max})$, contradicting the local expansion in Lemma 2.

Therefore, $\mathcal{Z}$ cannot have an accumulation point in $[0, T_{\max})$, which guarantees that $\mathcal{Z}$ is a discrete set containing only isolated points. As a result, exact layer balance $\mathbf{D}(t) = \mathbf{0}$ can never persist over any non-degenerate interval $[t_a, t_b] \subset [0, T_{\max})$ with $t_a < t_b$, completing the proof. □

Theorem 1 demonstrates that Feedback Alignment and standard gradient descent diverge fundamentally at the level of layer-wise invariants. We verify this empirically in the next section.

4 Empirical results

To validate our theoretical findings, we track the *layer balance gap* $\mathbf{D}_{\ell,\ell+1}(t) := \|W_{\ell+1}^\top(t)W_{\ell+1}(t) - W_\ell(t)W_\ell^\top(t)\|_F$ between adjacent weight matrices at epoch t, where $\|\cdot\|_F$ denotes the Frobenius norm. Across linear networks on synthetic matrix detection tasks and deep non-linear networks on CIFAR-10 and MNIST image classification tasks, we analyze how biologically plausible algorithms `RFA and DFA` breaks initial balancedness condition and how these dynamics dictate performance gains.

4.1 Empirical Validation on Matrix Detection

We evaluate our theoretical predictions in a controlled setup using a 3-layer linear network on the synthetic matrix detection task. The task is to learn a 50×100 matrix. All models are trained via

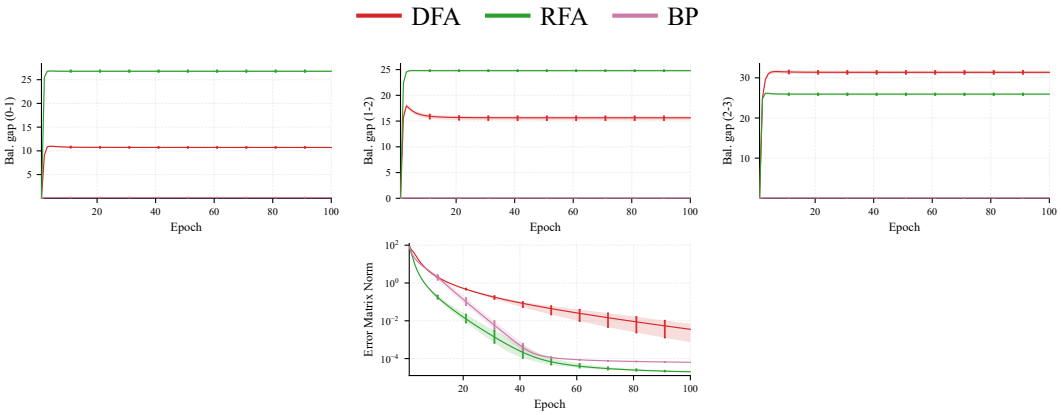

Figure 1: **Layer Balance and Error Matrix Norms in Matrix Detection.** Top row: layer discrepancy gaps $\mathbf{D}_{\ell,\ell+1}(t)$ (`Bal.gap`) across all transitions. Bottom row: error matrix norms. Solid lines and shaded bands represent the seed mean and min–max range over training epochs (x-axis). While `BP` maintains its initial zero-balanced state throughout training, `RFA` and `DFA` progressively diverge, confirming our theoretical prediction.

SGD on mean squared error (MSE) loss with a learning rate of $\eta = 0.1$, with results averaged across 7 independent random seeds. As shown in Figure 1 while standard `BP` maintains its initial zero-balance throughout training, the biologically plausible algorithms (`RFA` and `DFA`) rapidly break δ-balancedness, leading to pronounced layer imbalance. Nevertheless, all three methods successfully learn, with `RFA` achieving slightly superior optimization performance in the first 100 epochs.

4.2 Empirical Validation on CIFAR-10

To evaluate how weight balance behaves in non-linear settings, we train a 4-layer fully connected network with layer widths $[1024, 512, 256, 128]$ on CIFAR-10 for 50 epochs using SGD ($\eta = 0.1$), tanh activations, and cross-entropy loss, reporting results averaged over 5 independent random seeds. To isolate the effects of weight norm versus initial structural alignment, we evaluate two rescaled variants alongside standard `Xavier` and `Arora Balanced` initializations: (i) `Arora Balanced + XavNorm`, where weight layers are individually rescaled after a balanced initialization to match Xavier variance, and (ii) `Arora Balanced (std:1000)`, where weights are sampled with a high standard deviation. Crucially, while all balanced variants differ purely in scale, they preserve initial singular vector coupling across adjacent layers—a property absent under standard `Xavier` initialization.

Figure 2 illustrates the training loss trajectories and layer balance gaps $\mathbf{D}_{\ell,\ell+1}(t)$ across feedback matrix scales $\gamma \in \{0.025, 0.050, 0.1\}$.The empirical results reveal a clear scale-dependent trade-off governed by the feedback scale γ:

- **Low Feedback Scale ($\gamma = 0.025, 0.05$):** `Arora Balanced` initialization achieves superior optimization speed and lower final training loss relative to standard `Xavier` baselines.
- **High Feedback Scale ($\gamma = 0.1$):** Standard `Xavier` initialization exhibits greater numerical stability and outperforms balanced variants as γ increases.
- **Decay of Structural Coupling:** Table 1 reports the average cosine similarity between the left singular vectors of W_ℓ and the right singular vectors of $W_{\ell+1}$. While `Arora Balanced` initialization gradually loses its initial inter-layer coupling over the course of training, its alignment remains consistently higher than that of `Xavier` across all layers after 50 epochs.
- **Role of Weight Scaling:** Both `Arora Balanced + XavNorm` and `Arora Balanced (std:1000)` achieve optimization trajectories closely tracking unscaled `Arora Balanced`, suggesting that initial singular vector alignment could be driving the early feature learning.

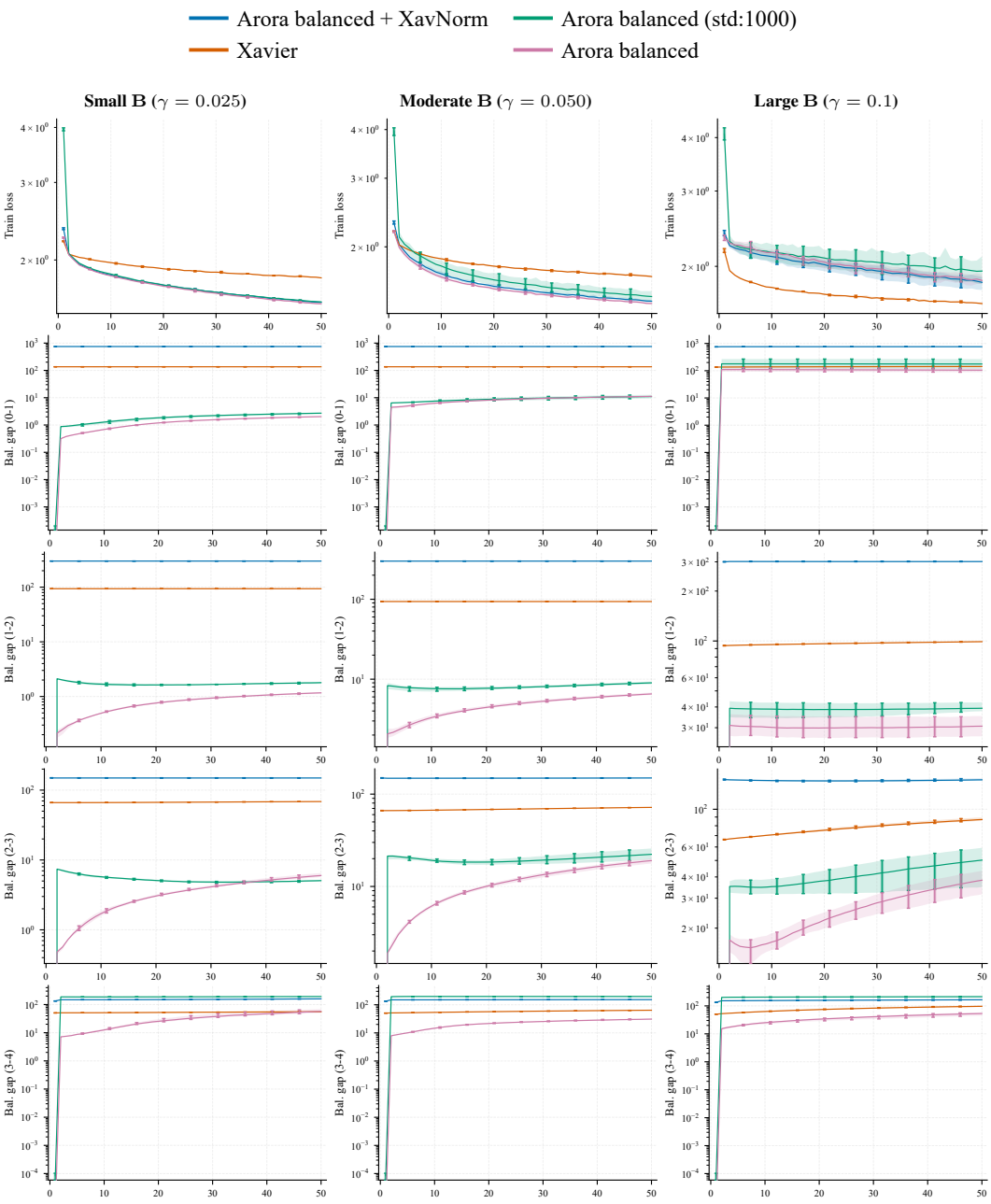

Figure 2: **Layer balance dynamics and Training Loss of RFA on CIFAR-10.** The top row depicts training loss, followed by layer balance gaps $\mathbf{D}_{\ell,\ell+1}(t)$(`Bal.gap`) for all transitions. x-axis represents the epoch. Solid lines denote the mean across random seeds over training epochs (x-axis), with shaded bands indicating the min–max range.. `Balanced` initialization yields superior training performance at smaller feedback scales ($\gamma = 0.025$), whereas `Xavier` initialization performs better at larger feedback scales ($\gamma = 0.1$). Although `Arora balanced + XavNorm` initialization maintains a high `Bal.gap` due to per-layer norm adjustment, it performs like `Arora Balanced` initialization. The initialization with `Arora Balanced(std:1000)` despite its high initial norm, also has a performance close to `Balanced` initialization.

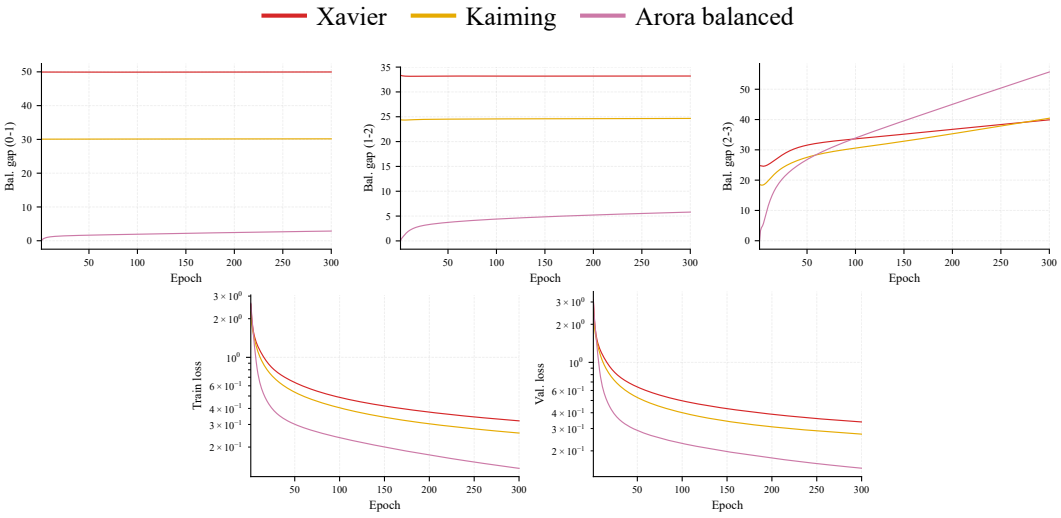

Figure 3: **Layer Balance and for MNIST.** Top row: layer discrepancy gaps $\mathbf{D}_{\ell,\ell+1}(t)$ (`Bal.gap`) across all transitions. Bottom row: Train and Validation Loss. We observe that `Arora balanced` performs better that other unstructured initializations.

Collectively, these findings demonstrate that balanced initialization accelerates training when paired with an appropriately calibrated feedback scale γ. We conjecture that strong initial structural coupling positions the network in a favorable regime of the loss landscape, facilitating faster training. A formal theoretical analysis of the convergence rates under balanced initialization, remains an important avenue for future work.

4.3 Empirical Validation on MNIST

To test whether the benefits of balanced initialization generalize to other datasets, we extend our evaluation to MNIST. Specifically, we train a 3-layer fully connected network with layer widths $[128, 64, 32]$, under RFA for 300 epochs using SGD ($\eta = 0.1$), tanh activations, and cross entropy loss. We use an 80/20 split of the training data. Mirroring our observations on CIFAR-10, structured `Arora Balanced` initialization consistently outperforms standard unstructured baselines, such as `Kaiming` and `Xavier` as shown in figure 3

5 Discussion

We summarise our findings and outline future directions below.

Theory. We proved that balancedness is not an invariant of Random Feedback Alignment: the discrepancy matrix is real-analytic with a nonzero quadratic term at zero initialization, so exact balance can hold only at isolated time points. Our analysis covers two-layer linear networks with zero initialization. Since the proof only requires the vector field to be real-analytic, not linear, we conjecture the result extends to networks with real-analytic nonlinearities.

Empirics. We showed that balanced initialization still improves training under RFA at small feedback scales, despite the invariant being broken. Our diagnostic variants ruled out balance gap and weight scale as sole explanations, but the exact source of the gain remains unclear. We conjecture that balanced initialization places training in a favorable region of the loss landscape.

Future Directions. Understanding how RFA perturbs singular vectors during training may explain the initialization-dependent gains. Extending the invariant analysis to Direct Feedback Alignment is another natural step. Since BP and RFA share the same underlying loss landscape, progress on these questions would shed light on the general loss landscape, not just RFA.

AI USE STATEMENT

In this work, we used generative AI tools to provide critical ingredients for proving mathematical claims, assist in the writing of proofs, interpret results, implement methods, and design or provide feedback on research methodology or experiments. We have not used generative AI tools to formulate mathematical claims, generate synthetic data sets, help develop theoretical models or conceptual frameworks, propose or refine hypotheses, support qualitative and thematic data analysis, clean and reformat datasets, or assist with translation; the rest of the tasks are not applicable to this work. Additionally, we used generative AI tools to brainstorm, summarize or analyse existing literature, identify relevant literature, format references, suggest a structure for a research paper, propose a title or keywords for a research paper, draft parts of a research paper, and source or search for information. We have reviewed all AI-assisted work. We corrected multiple critical math errors and hallucinations made by the LLM while proving various parts of the theoretical section, and we verified the corrected claims and proofs against standard mathematical texts and established results. 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 does not involve human subjects, personal data, or practices with potential ethical concerns. The datasets used are standard public benchmarks (CIFAR-10, MNIST), and the research is theoretical and empirical in nature. We therefore do not identify any ethical issues.

REPRODUCIBILITY STATEMENT

For the theoretical results, complete proofs of all claims, including the necessary lemmas and standard analytic results, are provided in Appendix, with standard references cited therein. For the empirical results, we describe the architectures, hyperparameters, and training details in Section 4. Our RFA and DFA implementations are based on the publicly available code from the paper (Sanfiz & Akrout, 2021). We first developed a custom implementation of RFA and DFA, then validated it against the BioTorch implementation to ensure the results match. Seeds were taken to be random and before each run we verified that the layer weights and feedback weights were matched and consistent across baselines.

REFERENCES

Mohamed Akrout, Collin Wilson, Peter Humphreys, Timothy Lillicrap, and Douglas B. Tweed. Deep learning without weight transport. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Vladimir I. Arnold. *Ordinary Differential Equations*. Springer-Verlag, Berlin, Heidelberg, 3rd edition, 1992.

Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. A convergence analysis of gradient descent for deep linear neural networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=SkMQg3C5K7.

Randolph E. Bank and Ronald L. Graham. *Numerical Analysis and Ordinary Differential Equations*. Academic Press, New York, 1991.

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. *Advances in Neural Information Processing Systems*, 31, 2018.

Gauthier Boeshertz, Razvan Pascanu, and Claudia Clopath. Overcoming rank collapse in feedback alignment. 2026. URL https://api.semanticscholar.org/CorpusID:289132792.

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.

Dominique Chu and Florian Bacho. Random feedback alignment algorithms to train neural networks: why do they align? *Machine Learning: Science and Technology*, 5(025023), 2024. doi: 10.1088/2632-2153/ad3ee5.

Francis Crick. The recent excitement about neural networks. *Nature*, 337(6203):129–132, 1989.

Wojciech Marian Czarnecki, Simon Osindero, Max Jaderberg, Grzegorz Świrszcz, and Razvan Pascanu. Understanding synthetic gradients and decoupled neural interfaces. In *Proceedings of the International Conference on Machine Learning*, volume 70, pp. 904–912, 2017.

Simon S Du, Wei Hu, and Jason Lee. Algorithmic Regularization in Learning Deep Homogeneous Models: Layers Are Automatically Balanced. In *Neural Information Processing Systems*, pp. 384–395, 2018. URL https://mlanthology.org/neurips/2018/du2018neurips-algorithmic/.

Jeffrey L. Elman, Elizabeth A. Bates, Mark H. Johnson, Annette Karmiloff-Smith, Domenico Parisi, and Kim Plunkett. *Rethinking Innateness: A Connectionist Perspective on Development*. MIT Press, Cambridge, MA, 1996.

Charlotte Frenkel, Martin Lefebvre, and David Bol. Learning without feedback: Direct random target projection as a feedback-alignment algorithm with layerwise feedforward training. 2019.

Stephen Grossberg. Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science*, 11(1):23–63, 1987.

Ernst Hairer, Syvert Paul Nørsett, and Gerhard Wanner. *Solving Ordinary Differential Equations I: Nonstiff Problems*, volume 8 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, Heidelberg, 2nd edition, 1993.

Maher Hanut and Jonathan Kadmon. Training large neural networks with low-dimensional error feedback. *arXiv preprint arXiv:2502.20580*, 2025.

Max Jaderberg, Wojciech Marian Czarnecki, Simon Osindero, Oriol Vinyals, Alex Graves, David Silver, and Koray Kavukcuoglu. Decoupled neural interfaces using synthetic gradients. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 1627–1635, 2017.

Daniel Kunin, Aran Nayebi, Javier Sagastuy-Brena, Surya Ganguli, Jonathan Bloom, and Daniel Yamins. Two routes to scalable credit assignment without weight symmetry. In *International Conference on Machine Learning*, pp. 5511–5521. PMLR, 2020.

Julien Launay, Iacopo Poli, and Florent Krzakala. Principled training of neural networks with direct feedback alignment. *arXiv preprint arXiv:1906.04554*, 2019.

Julien Launay, Iacopo Poli, François Boniface, and Florent Krzakala. Direct feedback alignment scales to modern deep learning tasks and architectures. *arXiv preprint arXiv:2006.12878*, 2020.

Yunseok Lee and Seulki Lee. Enabling fine-tuning of direct feedback alignment via feedback-weight matching. In *The Fourteenth International Conference on Learning Representations*, 2026.

Qianli Liao, Joel Leibo, and Tomaso Poggio. How important is weight symmetry in backpropagation? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

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 (13276), 2016.

Adam H. Marblestone, Greg Wayne, and Konrad P. Kording. Toward an integration of deep learning and neuroscience. *Frontiers in Computational Neuroscience*, 10:94, 2016.

David A. Medler. A brief history of connectionism. *Neural Computing Surveys*, 1:61–101, 1998.

Theodore H. Moskovitz, Ashok Litwin-Kumar, and L. F. Abbott. Feedback alignment in deep convolutional networks. *arXiv preprint arXiv:1812.06488*, 2018.

Arild Nøkland. Direct feedback alignment provides learning in deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 1037–1045, 2016.

Maria Refinetti, Stéphane d'Ascoli, Ruben Ohana, and Sebastian Goldt. Align, then memorise: The dynamics of learning with feedback alignment. In *International Conference on Machine Learning*, pp. 8925–8935. PMLR, 2021.

Zachary Robertson and Sanmi Koyejo. Layer-Wise Feedback Alignment Is Conserved in Deep Neural Networks. In *ICML 2023 Workshops: LLW*, 2023. URL https://mlanthology.org/icmlw/2023/robertson2023icmlw-layerwise/.

Zachary Robertson and Sanmi Koyejo. Implicit Regularization in Feedback Alignment Learning Mechanisms for Neural Networks. In *International Conference on Machine Learning*, volume 235, pp. 42601–42619, 2024. URL https://mlanthology.org/icml/2024/robertson2024icml-implicit/.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986a. doi: 10.1038/323533a0.

David E. Rumelhart, James L. McClelland, and the PDP Research Group. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 2: Psychological and Biological Models*. MIT Press, Cambridge, Massachusetts, 1986b.

Albert Jiménez Sanfiz and Mohamed Akrout. Benchmarking the accuracy and robustness of feedback alignment algorithms. *CoRR*, abs/2108.13446, 2021. URL https://arxiv.org/abs/2108.13446.

Ganlin Song, Ruitu Xu, and John D. Lafferty. Convergence and Alignment of Gradient Descent with Random Backpropagation Weights. In *Neural Information Processing Systems*, 2021. URL https://mlanthology.org/neurips/2021/song2021neurips-convergence/.

Gerald Teschl. *Ordinary Differential Equations and Dynamical Systems*, volume 140 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, Rhode Island, 2012.

Will Xiao, Honglin Chen, Qianli Liao, and Tomaso Poggio. Biologically-plausible learning algorithms can scale to large datasets. *arXiv preprint arXiv:1811.03567*, 2018.

A KEY RESULTS FROM ANALYSIS

For completeness, this section collects standard foundational results from the classical theory of ordinary differential equations (ODEs) and real analysis. Comprehensive treatments and canonical proofs for these analytic properties can be found in foundational texts such as Hairer et al. (1993), Arnold (1992), and Teschl (2012).

The Cauchy-Kovalevski theorem guarantees that a unique analytic solution exists for a partial differential equation (or system) with analytic initial conditions around a given point.

Theorem 2 (Cauchy–Kovalevski Theorem for ODEs)**.** *Let $\Omega \subseteq \mathbb{R}^d$ be an open neighborhood containing $\mathbf{y}_0$. Suppose a vector field $\mathbf{f} : \Omega \to \mathbb{R}^d$ is real-analytic on Ω. Then the initial-value problem*

$$\dot{\mathbf{y}}(t) = \mathbf{f}(\mathbf{y}(t)), \quad \mathbf{y}(0) = \mathbf{y}_0 \tag{23}$$

possesses a unique solution $\mathbf{y}(t)$ that is real-analytic on some open interval $(-\epsilon, \epsilon)$ for $\epsilon > 0$.

To analyze the global properties of trajectories generated by such real-analytic vector fields, we additionally leverage the rigidity of analytic functions under the classical Identity Theorem.

Theorem 3 (Identity Theorem for Real-Analytic Functions)**.** *Let $g : I \to \mathbb{R}$ be a real-analytic function on an open connected interval $I \subseteq \mathbb{R}$. If the set of zeros $\{t \in I \mid g(t) = 0\}$ possesses an accumulation point in I, then $g(t) \equiv 0$ identically across I.*

B DISCRETE TO CONTINUOUS FRAMEWORK

This section unifies the discrete numerical approximations produced by Euler iterates with continuous differential equation flows. We begin by establishing the geometric compactness of the flow domain, move through the foundational local estimates detailed in Section 1.7 of Bank and Graham Bank & Graham (1991), and conclude with global convergence up to $T_{\max}$.

B.1 DOMAIN COMPACTNESS

To establish existence and convergence properties, we first prove that the domain of our continuous flow forms a compact subset of Euclidean space.

Lemma 3 (Compactness of the Flow Domain)**.** *Let $K = \{\boldsymbol{y} \in \mathbb{R}^{2mn} : \|\boldsymbol{y}\| \leq b\}$ represent the closed ball of radius $b > 0$ centered at the origin in $\mathbb{R}^{2mn}$, and let $[0, \tau] \subset \mathbb{R}$ be a closed time interval. The cylinder $D = [0, \tau] \times K$ is a compact subset of Euclidean space $\mathbb{R} \times \mathbb{R}^{2mn} \cong \mathbb{R}^{2mn+1}$.*

Proof. By the Heine–Borel theorem, a subset of Euclidean space is compact if and only if it is both bounded and closed.

- **Boundedness:** The time interval $[0, \tau]$ is strictly bounded by τ. Furthermore, the spatial domain K is a closed ball of radius b, meaning that $\|\boldsymbol{y}\| \leq b$ for all $\boldsymbol{y} \in K$. Because both components of the Cartesian product $D = [0, \tau] \times K$ are bounded in their respective spaces, every point $(t, \boldsymbol{y}) \in D$ satisfies

  $$\|(t, \boldsymbol{y})\|^2 = |t|^2 + \|\boldsymbol{y}\|^2 \leq \tau^2 + b^2 < \infty.$$

  Therefore, the entire cylinder D is bounded.

- **Closedness:** The interval $[0, \tau]$ contains all of its limit points, making it a closed set in $\mathbb{R}$. The spatial set K is defined by the inequality $\|\boldsymbol{y}\| \leq b$, which guarantees that it contains its boundary and forms a closed set in $\mathbb{R}^{2mn}$. The Cartesian product of two closed sets is closed in the product topology. Therefore, D is closed in $\mathbb{R}^{2mn+1}$.

Because D is both closed and bounded, it is compact. This compactness guarantees that every continuous vector field $\boldsymbol{f}(\boldsymbol{y})$ defined on D attains a finite maximum, yielding the uniform bound required for existence and uniqueness of the continuous flow. □

B.2 Classical Euler Polygon Theory

For complete derivations, detailed stability bounds, and proofs regarding explicit local estimates, refer directly to Section 1.7 of Bank and Graham Bank & Graham (1991).

We consider the initial value problem

$$\boldsymbol{y}'(x) = \boldsymbol{f}(x, \boldsymbol{y}), \quad \boldsymbol{y}(x_0) = \boldsymbol{y}_0, \tag{24}$$

where $\boldsymbol{f}$ is defined on an open set $U \subseteq \mathbb{R} \times \mathbb{R}^m$. Vector norms replace absolute values where appropriate.

Definition 2 (Euler polygon; see (Bank & Graham, 1991, Section 1.7))**.** Given a mesh subdivision $x_0 < x_1 < \cdots < x_n = X$ with step sizes $h_i = x_{i+1} - x_i$ and maximal step $|h| = \max_i h_i$, the explicit Euler scheme computes

$$\boldsymbol{y}_{i+1} = \boldsymbol{y}_i + h_i \boldsymbol{f}(x_i, \boldsymbol{y}_i), \quad i = 0, \ldots, n-1. \tag{25}$$

The Euler polygon $\boldsymbol{y}_h(x)$ is the piecewise linear interpolant:

$$\boldsymbol{y}_h(x) = \boldsymbol{y}_i + (x - x_i)\boldsymbol{f}(x_i, \boldsymbol{y}_i), \quad x_i \le x \le x_{i+1}. \tag{26}$$

The core properties established in Section 1.7 of Bank and Graham Bank & Graham (1991) are summarized below:

Lemma 4 (Cylinder containment and local consistency)**.** *Let $D = [x_0, X] \times \overline{B}(\boldsymbol{y}_0, b)$ be a compact cylinder on which $\|\boldsymbol{f}(x, \boldsymbol{y})\| \le A$. If $X - x_0 \le b/A$, then for every subdivision the numerical solution remains in D, and*

$$\|\boldsymbol{y}_h(x) - \boldsymbol{y}_0\| \le A|x - x_0|,$$
$$\|\boldsymbol{y}_h(x) - \boldsymbol{y}_0 - (x - x_0)\boldsymbol{f}(x_0, \boldsymbol{y}_0)\| \le \epsilon|x - x_0|$$

whenever $\|\boldsymbol{f}(x, \boldsymbol{y}) - \boldsymbol{f}(x_0, \boldsymbol{y}_0)\| \le \epsilon$ on D.

Lemma 5 (Sensitivity to initial values)**.** *Let $\boldsymbol{y}_h$ and $\boldsymbol{z}_h$ be two Euler polygons on the same subdivision h, starting from $\boldsymbol{y}_0$ and $\boldsymbol{z}_0$ respectively. If $\left\| \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{y}}(x, \boldsymbol{y}) \right\| \le L$ in a convex region that contains the graphs of both polygons for $x_0 \le x \le X$, then*

$$\|\boldsymbol{z}_h(x) - \boldsymbol{y}_h(x)\| \le e^{L(x - x_0)} \|\boldsymbol{z}_0 - \boldsymbol{y}_0\|. \tag{27}$$

Theorem 4 (Cauchy–Lipschitz via Euler polygons)**.** *Assume $\boldsymbol{f}$ is continuous on $D = [x_0, X] \times \overline{B}(\boldsymbol{y}_0, b)$. If $\|\boldsymbol{f}\| \le A$ on D, and the Lipschitz condition $\|\boldsymbol{f}(x, \boldsymbol{y}_1) - \boldsymbol{f}(x, \boldsymbol{y}_2)\| \le L\|\boldsymbol{y}_1 - \boldsymbol{y}_2\|$ holds on D, then if $X - x_0 \le b/A$, as $|h| \to 0$ the Euler polygons converge uniformly to a function $\boldsymbol{\phi}(x)$ which is the unique solution on $[x_0, X]$.*

Theorem 5 (Extension up to the boundary)**.** *Let $U \subseteq \mathbb{R}^{m+1}$ be open and $\boldsymbol{f}, \partial \boldsymbol{f}/\partial \boldsymbol{y}$ continuous on U. Then for every $(x_0, \boldsymbol{y}_0) \in U$ there exists a unique solution that extends (forward and backward) up to the boundary of U.*

B.3 Global Convergence up to $T_{\max}$

Theorem 6 (Global convergence up to $T_{\max}$)**.** *Let the hypotheses of Theorem 5 hold and let $\boldsymbol{y}(x)$ be the unique solution on the maximal forward interval $[x_0, T_{\max})$. For every T with $x_0 < T < T_{\max}$, the Euler polygons $\boldsymbol{y}_h(x)$ with maximum step size $|h| \to 0$ converge uniformly to $\boldsymbol{y}(x)$ on $[x_0, T]$.*

Proof. Fix $T < T_{\max}$. We prove the convergence by an inductive "bridging" argument. The key is to show that once convergence holds up to a point $X \in [x_0, T)$, it can be extended a little further to some $X' > X$, and by repeating finitely many times we reach T.

The single bridge. Assume that for some $X \in [x_0, T)$ we already have uniform convergence of $\boldsymbol{y}_h$ to $\boldsymbol{y}$ on $[x_0, X]$. Since $(X, \boldsymbol{y}(X)) \in U$ and U is open, we can choose a closed cylinder centred at the exact solution that lies entirely in U. Concretely, there exist $b > 0$ and $X_1 > X$ (with $X_1 \le T$) such that the compact set

$$D := [X, X_1] \times \overline{B}(\boldsymbol{y}(X), b)$$

is contained in U, and on D we have the bounds

$$\|\boldsymbol{f}(x,\boldsymbol{y})\| \le A, \qquad \left\|\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{y}}(x,\boldsymbol{y})\right\| \le L,$$

with $X_1 - X \le b/A$. (The bound on $\partial \boldsymbol{f}/\partial \boldsymbol{y}$ implies the Lipschitz condition on the convex set D.) Such D exists because we can first pick a small closed ball around $(X, \boldsymbol{y}(X))$ inside U, then fit an appropriate cylinder inside that ball, and finally shrink the cylinder further to satisfy the length condition $X_1 - X \le b/A$ if necessary.

Now let $\epsilon > 0$ be arbitrary.

Step 1: Shift the centre to the numerical value. Because $\boldsymbol{y}_h(X) \to \boldsymbol{y}(X)$ as $|h| \to 0$, there exists $\delta_0 > 0$ such that for all subdivisions with $|h| < \delta_0$,

$$\|\boldsymbol{y}_h(X) - \boldsymbol{y}(X)\| < \frac{b}{2}.$$

Define a smaller cylinder centred at $\boldsymbol{y}_h(X)$:

$$D' := [X, X'] \times \overline{B}\big(\boldsymbol{y}_h(X),\, \tfrac{b}{2}\big), \qquad \text{where } X' := X + \min\Big(X_1 - X,\ \frac{b/2}{A}\Big).$$

For any $(x, \boldsymbol{z}) \in D'$ we have

$$\|\boldsymbol{z} - \boldsymbol{y}(X)\| \le \|\boldsymbol{z} - \boldsymbol{y}_h(X)\| + \|\boldsymbol{y}_h(X) - \boldsymbol{y}(X)\| < \frac{b}{2} + \frac{b}{2} = b,$$

so $D' \subset D$. In particular, both cylinders are convex and the Lipschitz condition holds on them.

Step 2: The Euler polygon remains inside D'. Apply Lemma 4 to the cylinder D' with initial point $(X, \boldsymbol{y}_h(X))$. Because $X' - X \le \frac{b/2}{A}$, the lemma guarantees that the Euler polygon $\boldsymbol{y}_h(x)$ on $[X, X']$ stays entirely in D' (hence in D).

Step 3: Auxiliary polygon starting from the exact value. On $[X, X']$, construct an auxiliary Euler polygon $\boldsymbol{z}_h(x)$ that uses exactly the same subdivision as $\boldsymbol{y}_h$ but with initial value $\boldsymbol{z}_h(X) = \boldsymbol{y}(X)$. Since $\boldsymbol{y}(X)$ is the centre of the original cylinder D and $X' - X \le b/A$, Theorem 4 (applied to D with initial condition $\boldsymbol{y}(X)$) tells us that for sufficiently small $|h|$,

$$\sup_{x \in [X, X']} \|\boldsymbol{z}_h(x) - \boldsymbol{y}(x)\| < \frac{\epsilon}{2}. \tag{1}$$

Step 4: Compare the two polygons. Both $\boldsymbol{y}_h$ and $\boldsymbol{z}_h$ are Euler polygons on $[X, X']$ with the same subdivision. Their graphs lie in the convex set D (the auxiliary polygon stays close to the exact solution, which is in D; $\boldsymbol{y}_h$ stays in $D' \subset D$). Therefore Lemma 5 applies and yields

$$\|\boldsymbol{y}_h(x) - \boldsymbol{z}_h(x)\| \le e^{L(x-X)} \|\boldsymbol{y}_h(X) - \boldsymbol{y}(X)\|, \qquad x \in [X, X']. \tag{2}$$

Choose $\delta_1 > 0$ such that for $|h| < \delta_1$,

$$\|\boldsymbol{y}_h(X) - \boldsymbol{y}(X)\| < \frac{\epsilon}{2}\, e^{-L(T - x_0)}.$$

Since $x \le T$, the factor $e^{L(x-X)} \le e^{L(T-x_0)}$, and (2) gives

$$\|\boldsymbol{y}_h(x) - \boldsymbol{z}_h(x)\| < \frac{\epsilon}{2}. \tag{3}$$

Step 5: Error on the extended interval. Combining (1) and (3) via the triangle inequality,

$$\sup_{x \in [X, X']} \|\boldsymbol{y}_h(x) - \boldsymbol{y}(x)\| \le \sup \|\boldsymbol{y}_h - \boldsymbol{z}_h\| + \sup \|\boldsymbol{z}_h - \boldsymbol{y}\| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,$$

provided $|h| < \min(\delta_0, \delta_1,$ the threshold for $(1))$. Thus $\boldsymbol{y}_h \to \boldsymbol{y}$ uniformly on $[X, X']$.

If $X' < X_1$, we can repeat the same argument on the next subinterval $[X', X_1]$ (now with the new exact point $(X', \boldsymbol{y}(X'))$, which again lies in some cylinder inside U). After another repetition, we cover the whole $[X, X_1]$.

From x_0 to T. Start at $X = x_0$, where convergence holds trivially because $\boldsymbol{y}_h(x_0) = \boldsymbol{y}_0 = \boldsymbol{y}(x_0)$. Apply the bridge argument repeatedly: choose a first cylinder $D_0 = [x_0, X^{(1)}] \times \overline{B}(\boldsymbol{y}_0, b_0)$ given by local existence, obtain convergence up to $X^{(1)}$; then take $(X^{(1)}, \boldsymbol{y}(X^{(1)}))$ as the new base point and repeat. Because $[x_0, T]$ is compact, we can select finitely many such bridges that reach T. After the last step we have uniform convergence on the whole $[x_0, T]$. □

The crucial insight rests on the numerical value $\boldsymbol{y}_h(X)$ growing arbitrarily close to the exact trajectory value $\boldsymbol{y}(X)$ as $|h| \to 0$. This proximity allows us to embed a smaller cylinder D' around $\boldsymbol{y}_h(X)$ within the safe cylinder D. Lemma 4 retains subsequent numerical iterates inside D', while the continuous solution remains in D by design. Thus, all estimates occur inside the convex Lipschitz domain D, where Lemma 5 compares the respective polygons. Repeating this single-bridge extension finitely many times covers any compact subinterval of $[x_0, T_{\max})$.