How and When Random Feedback Works:
A Case Study of Low-Rank Matrix Factorization

Shivam Garg
Stanford University
shivamg@cs.stanford.edu

Santosh S. Vempala
Georgia Tech
vempala@gatech.edu

Abstract

The success of gradient descent in ML and especially for learning neural networks is remarkable and robust. In the context of how the brain learns, one aspect of gradient descent that appears biologically difficult to realize (if not implausible) is that its updates rely on feedback from later layers to earlier layers through the same connections. Such bidirected links are relatively few in brain networks, and even when reciprocal connections exist, they may not be equi-weighted. Random Feedback Alignment (Lillicrap et al., 2016), where the backward weights are random and fixed, has been proposed as a bio-plausible alternative and found to be effective empirically. We investigate how and when feedback alignment (FA) works, focusing on one of the most basic problems with layered structure — low-rank matrix factorization. In this problem, given a matrix $Y_{n \times m}$, the goal is to find a low rank factorization $Z_{n \times r} W_{r \times m}$ that minimizes the error $\|ZW - Y\|_F$. Gradient descent solves this problem optimally. We show that FA finds the optimal solution when $r \geq \mathrm{rank}(Y)$. We also shed light on *how* FA works. It is observed empirically that the forward weight matrices and (random) feedback matrices come closer during FA updates. Our analysis rigorously derives this phenomenon and shows how it facilitates convergence of FA*, a closely related variant of FA. We also show that FA can be far from optimal when $r < \mathrm{rank}(Y)$. This is the first provable separation result between

gradient descent and FA. Moreover, the representations found by gradient descent and FA can be almost orthogonal even when their error $\|ZW - Y\|_F$ is approximately equal. As a corollary, these results also hold for training two-layer linear neural networks when the training input is isotropic, and the output is a linear function of the input.

1 INTRODUCTION

Information Processing in the brain is hierarchical, with multiple layers of neurons from perception to cognition, and learning is believed to be largely based on updates to synaptic weights. These weight updates depend on error information that may only be available in the downstream (higher-level) areas. An algorithmic challenge faced by the brain is the following: how to update the weights of earlier layers using the error information from later layers, despite local structural constraints? For example, in the visual cortex, the weight update to earlier layers — which detect low-level information such as edges in an image — may depend on higher-level information in the image that is available only after downstream processing.

In artificial neural networks, gradient descent via backpropagation (Rumelhart et al., 1986) has been a very successful method of making weight updates. However, it is unclear whether gradient descent is biologically plausible due to its non-local updates (Crick, 1989). In particular, the update to earlier layers involves feedback from later layers through backward weights that are transposed copies of the corresponding forward weights (see Fig. 1). This requires equi-weighted bidirectional links between neurons, which are rare in the brain. This issue was first identified by Grossberg (1987), who called it the weight transport problem.

Rather surprisingly, Lillicrap et al. (2016) found that

neural networks are able to learn even when the backward feedback weights are random and fixed, independent of the forward weights. This biologically plausible variant of gradient descent is known as *Feedback Alignment (FA)*. Feedback alignment and its variants (Nøkland, 2016) have been shown to be effective for many problems ranging from language modeling to neural view synthesis (Launay et al., 2020). At the same time, they do not match the performance of gradient descent for large-scale visual recognition problems (Bartunov et al., 2018; Moskovitz et al., 2018) such as ImageNet (Russakovsky et al., 2015).
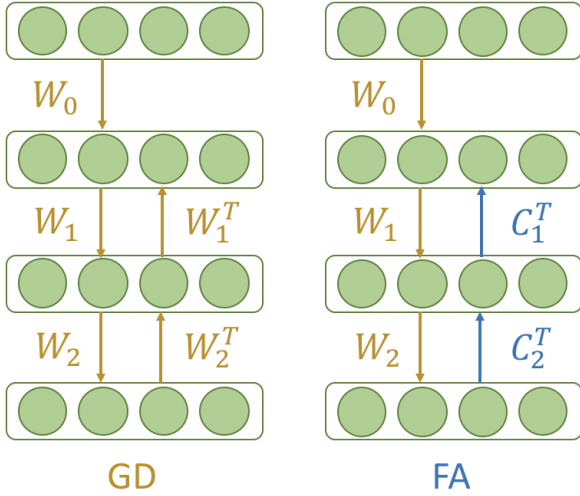

Figure 1: Gradient descent uses transpose of the forward weights for backward feedback while feedback alignment replaces them by fixed random weights.

These observations raise many questions: How and when does random feedback work? Is there any fundamental sense in which feedback alignment is inferior to gradient descent? How different are the representations found using feedback alignment and gradient descent? Alongside the biological motivation, these questions are also important for getting a better understanding of the landscape of possible optimization algorithms.

Problem formulation and contributions. In this paper, we investigate these questions by considering one of the most basic problems with layered structure — low-rank matrix factorization (Du et al., 2018; Valavi et al., 2020; Ye and Du, 2021). In this problem, given a matrix $Y_{n \times m}$, the goal is to find a low rank factorization $Z_{n \times r} W_{r \times m}$ that minimizes the error

$$\|Z_{n \times r} W_{r \times m} - Y_{n \times m}\|_F^2. \tag{1}$$

The gradient flow (GD) update (gradient descent with infinitesimally small step size) for this problem is given

by

$$\frac{dZ}{dt} = (Y - \hat{Y})W^T$$
$$\frac{dW}{dt} = Z^T(Y - \hat{Y}), \tag{2}$$

where $\hat{Y} = ZW$. From prior work (Bah et al., 2019, Theorem 39), we know that gradient flow starting from randomly initialized Z and W converges to the optimal solution almost surely. The layered structure and optimality of gradient flow makes low-rank matrix factorization an ideal candidate for understanding the performance of feedback alignment.

The feedback alignment (FA) update is given by

$$\frac{dZ}{dt} = (Y - \hat{Y})C^T$$
$$\frac{dW}{dt} = Z^T(Y - \hat{Y}) \tag{3}$$

Note that the only difference from gradient flow update is that the backward feedback weight W^T is replaced by C^T in the expression for $\frac{dZ}{dt}$. Here, C is some (possibly random) fixed matrix.

Empirically, it is observed that the backward feedback weights (C, in this case) and the forward weights (W) come closer during feedback alignment updates (Lillicrap et al., 2016). After the forward and backward weights are sufficiently aligned, the feedback alignment update is similar to the gradient flow update. This alignment between the forward weights and the backward feedback weights led to the name feedback alignment, and is considered to be the main reason for the effectiveness of this algorithm.

However, the phenomenon of alignment has turned out to be hard to establish rigorously. One reason behind this is that the alignment between forward and backward weights may not increase monotonically (see Example 1 for details). We observe that a small tweak to the feedback alignment update where W is updated optimally, leads to monotonically increasing alignment between C and W (for an appropriately defined notion of alignment). We call this version of feedback alignment FA*, and its updates are given by

$$\frac{dZ}{dt} = (Y - \hat{Y})C^T$$
$$W = (Z^T Z)^{-1} Z^T Y. \tag{4}$$

Notice that the only difference between FA and FA* is that W is chosen optimally (given Z and Y) in the FA* update, while it moves in the negative gradient direction in the FA update. The update to Z remains the same, and involves a fixed feedback matrix C.

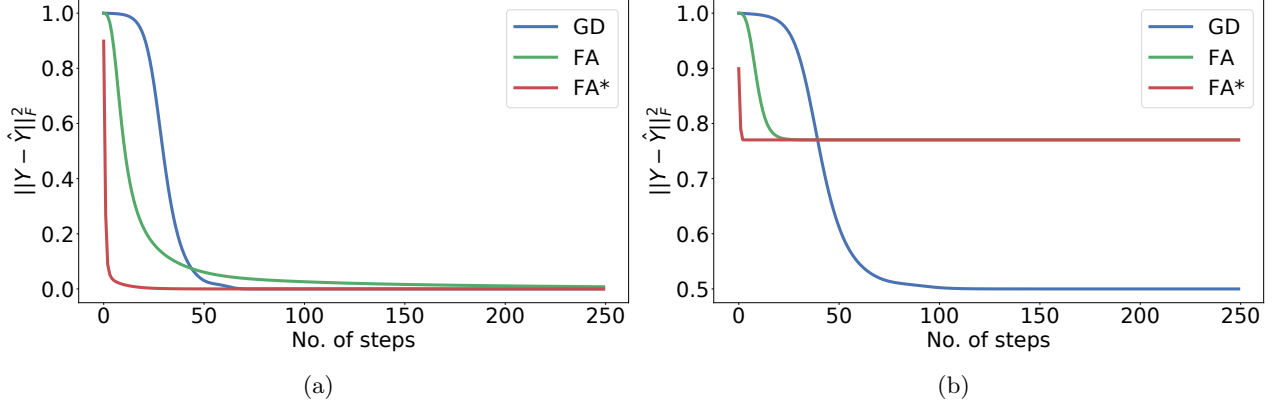We show that FA* initialized with an arbitrary full column rank Z converges to a stationary point where

Figure 2: (a) Feedback alignment (both FA (3) and FA* (4)) converge to the optimal solution when $r \geq rank(Y)$. In this plot, $n = m = 500$ and $r = rank(Y) = 50$. (b) Feedback alignment solution can be far from optimal when $r < rank(Y)$. In this plot, $n = m = rank(Y) = 500$ and $r = 50$. Gradient descent finds the optimal solution in both the cases.

$(Y - \hat{Y})C^T = 0$ (Theorem 1). Our analysis rigorously demonstrates the phenomenon of alignment, and sheds light on how it facilitates convergence (Section 3).

Convergence of FA to a stationary point has also been shown in past works (Baldi et al., 2018; Lillicrap et al., 2016). Baldi et al. (2018) proves convergence of feedback alignment for learning one-hidden layer neural networks with linear activation, starting from arbitrary initialization. For low-rank matrix factorization, this implies convergence of FA to a stationary point. However, as we discussed, alignment between forward and backward weights may not increase monotonically in FA. Due to this, these works are not able to say much about the dynamics of alignment. The main feature of our analysis is that by analyzing a slight variant of FA (FA*), we obtain a better understanding of the phenomenon of alignment and its implications for convergence.

After analysing how feedback alignment works, we shift our attention to the question of when it works. We characterize the solution $\hat{Y}$ at the stationary points of feedback alignment, when C is chosen randomly (Lemma 1). Building on this characterization, we show that feedback alignment finds the optimal solution when $r \geq rank(Y)$ (Theorem 2). However, it can be far from optimal when $r < rank(Y)$ (Theorem 3). To the best of our knowledge, this is the first provable separation result between gradient flow and feedback alignment (see Fig. 2 for an illustration).

Moreover, the representations found by feedback alignment and gradient flow are very different. We show that even when their errors $\|ZW - Y\|_F^2$ are approximately equal, the representations found (Z) can be almost orthogonal (Theorem 4).

Since the stationary point equations for FA and FA* are same, these results about suboptimality of feedback alignment and difference in representations apply to both versions of feedback alignment.

In summary, we give a comprehensive analysis of how and when feedback alignment works, focusing on the problem of low-rank matrix factorization. Here is a list of our contributions:

1. We prove convergence of feedback alignment (FA*) to a stationary point, shedding light on the dynamics of alignment and its implications for convergence (Section 3).

2. We show that feedback alignment (both FA and FA*) find the optimal solution when $r \geq rank(Y)$, but can be far from optimal when $r < rank(Y)$. This shows provable separation between feedback alignment and gradient flow (Section 4).

3. We characterize the representations found by feedback alignment (both FA and FA*), and show that they can be very different from the representations found by gradient flow, even when their errors are approximately equal (Section 4).

As a corollary, all our results also hold for training two-layer linear neural networks, assuming the training input is isotropic and the output is a linear function of the input (Section 5). We defer all proofs and simulation details to the appendix.

Notation. For any matrix M, $M(t)$ denotes its value at time t. We will not explicitly show t when it is clear from context. $\sigma_i(M)$ denotes the ith largest singular value of M. $M^{(i)}$ denotes the ith column of M. $\|M\|_F$

denotes the Frobenius norm of M and $\|v\|$ denotes the ℓ_2 norm of vector v.

2 RELATED WORK

Feedback alignment. Lillicrap et al. (2016) show convergence of feedback alignment dynamics for learning one-hidden layer neural networks with linear activation, starting from zero initialization. Baldi et al. (2018) generalize this result to arbitrary initialization, and also show convergence for linear neural networks of arbitrary depth when the input and all hidden layers are one dimensional.

In recent work, (Song et al., 2021) study feedback alignment for highly overparameterized one-hidden layer neural networks where the width of the hidden layer is much larger than the size of training set. This work builds on past work on Neural Tangent Kernels (Jacot et al., 2018), and shows that feedback alignment converges to a solution with zero training error. Contrary to the popular understanding of feedback alignment, they show that forward and backward weights may not align in this highly overparameterized regime. However, in the parameter regime typically encountered in practice, alignment is a robust phenomenon (Lillicrap et al., 2016).

Refinetti et al. (2021) obtain a set of ODEs that describe the progression of feedback alignment test error for neural networks in certain parameter regimes. Using simulations, and analysis of these ODEs at initialization, they argue that neural network training proceeds in two phases: the initial alignment phase where the forward and backward weights align with each other, followed by a memorization phase where learning happens. In Section 3, we show that while such a progression can take places in simple cases (see Example 1), in general, the dynamics are much more involved with highly interleaved phases. This paper also presents intuition about the behaviour of feedback alignment for deeper networks, and possible reasons for its poor performance with convolutional neural networks (CNNs).

The focus of our work is twofold: (i) understanding how feedback alignment works by studying the dynamics of alignment and its impact on convergence, (ii) understanding when feedback alignment works by contrasting its solution and representations with gradient descent. Our work complements the existing line of work on understanding feedback alignment.

Biologically plausible learning. Many algorithms have been proposed to address the weight transport problem (Lillicrap et al., 2020). Most of these algorithms either encourage alignment between forward and backward weights implicitly (Lillicrap et al., 2016; Nøkland, 2016; Moskovitz et al., 2018; Akrout et al., 2019), or learn weights that try to preserve information between adjacent layers (Bengio, 2014; Lee et al., 2015; Kunin et al., 2019, 2020). A parallel line of work studies how training algorithms can be implemented in the brain using spiking neurons without distinct inference (forward propagation) and training (backward propagation) phases (Xie and Seung, 2003; Bengio et al., 2017; Scellier and Bengio, 2017; Whittington and Bogacz, 2017; Guerguiev et al., 2017; Sacramento et al., 2018). More recent work more directly models plasticity and inhibition in the brain and shows that memorization and learning are emergent phenomena (Papadimitriou et al., 2020; Dabagia et al., 2021).

Building the mathematical foundation of such biologically plausible algorithms can lead to illuminating insights applicable to the brain as well as to the general theory of learning and optimization. Our work can be viewed as progress in this direction.

3 CONVERGENCE

In this section, we show that FA* (4) converges to a stationary point satisfying $(Y - \hat{Y})C^T = 0$, where $\hat{Y} = ZW$ and $W = (Z^T Z)^{-1} Z^T Y$.

Theorem 1. *Let $Z(0)$ be full column rank. For any $\epsilon > 0$ and*

$$T \geq \frac{24}{\epsilon} \left(\frac{\sigma_1\left(Y\right)\sigma_1\left(C\right)\sigma_1\left(Z(0)\right)^6 \sqrt{r\; min(m,n)}}{\sigma_r\left(Z(0)\right)^5} \right),$$

FA dynamics (4) satisfy*

$$\min_{t \leq T}\; \left\| (Y - \hat{Y}(t))C^T \right\|_F^2 \leq \epsilon.$$

Moreover,

$$\lim_{t \to \infty} \left\| (Y - \hat{Y}(t))C^T \right\|_F^2 = 0.$$

Note that the time for convergence of minimum of $\left\| (Y - \hat{Y}(t))C^T \right\|_F^2$ depends linearly on $\frac{1}{\epsilon}$. We describe the complete proof of Theorem 1 in Appendix A. We also describe the proof idea for the convergence result by Baldi et al. (2018) in Appendix E.

To understand our result, let's first discuss a toy example where $m = 1$ (recall Y is an $n \times m$ matrix).

Example 1. *Suppose we want to factorize $y_{n \times 1}$ as $\hat{y}_{n \times 1} = Z_{n \times r} w_{r \times 1}$, and we use $c_{r \times 1}$ for feedback. FA* update is given by*

$$\frac{dZ}{dt} = (y - \hat{y})c^T$$
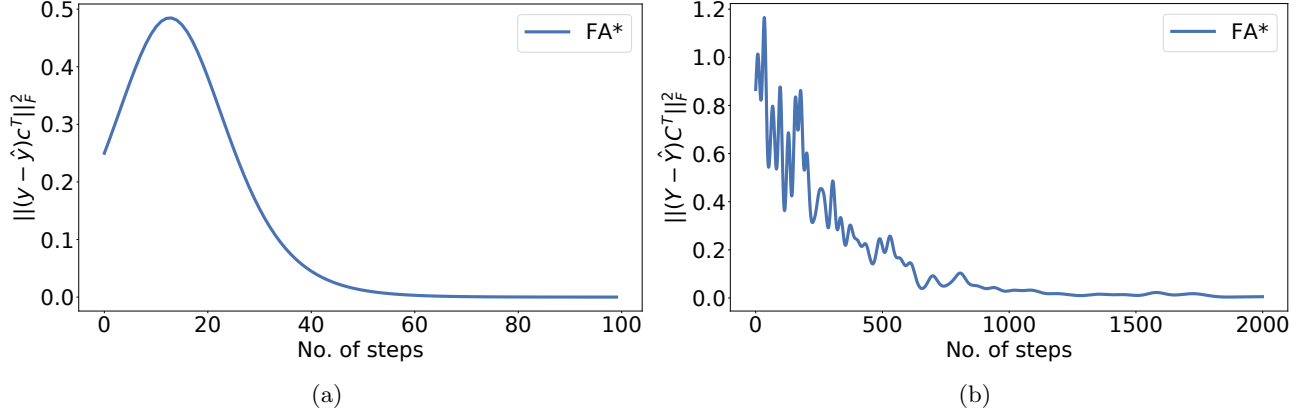$$w = (Z^T Z)^{-1} Z^T y.$$

Figure 3: (a)When y and $\hat{y}$ are column vectors, $\left\|(y-\hat{y})c^T\right\|_F^2$ increases monotonically in the beginning followed by monotonic decrease (see Example 1). (b) For general matrices Y and $\hat{Y}$, progression of $\left\|(Y-\hat{Y})C^T\right\|_F^2$ can be highly non-monotonic.

This gives

$$\frac{d \left\|(y-\hat{y})c^T\right\|_F^2}{dt} = -2\left\|y-\hat{y}\right\|^2 \left\|c\right\|^2 c^T w$$

$$\frac{d\ c^T w}{dt} = c^T(Z^T Z)^{-1}c \ \left\|y-\hat{y}\right\|^2.$$

Note that $\frac{d\ c^T w}{dt} \geq 0$. And $\frac{d \left\|(y-\hat{y})c^T\right\|_F^2}{dt} \geq 0$ only when $c^T w < 0$. So $c^T w$ increases with time (when $\left\|y-\hat{y}\right\|^2 > 0$). $\left\|y-\hat{y}\right\|^2\left\|c\right\|^2$ increases in the beginning if $c^T w < 0$, but it starts decreasing once $c^T w > 0$, and eventually goes to 0 (see Fig. 3a). This shows how alignment of c and w (measured by $c^T w$) facilitates convergence.

The alignment between c and w increases monotonically in FA* dynamics. This is not true in FA dynamics. FA update is given by

$$\frac{dZ}{dt} = (y-\hat{y})c^T$$

$$\frac{dw}{dt} = Z^T(y-\hat{y}).$$

This gives

$$\frac{d\ c^T w}{dt} = c^T Z^T(y-\hat{y}).$$

Suppose $Z(0) = -yc^T$ and $w(0) = 0$. In this case, at $t = 0$, $\frac{d\ c^T w}{dt} = -\left\|c\right\|^2\left\|y\right\|^2 < 0$. Therefore, alignment between c and w can also decrease in FA dynamics. This makes FA* more suitable to understand the dynamics of alignment and its implications for convergence.

In this example with $m = 1$, we saw that the loss $\left\|(Y-\hat{Y})C^T\right\|_F^2$ has an initial phase in which it can increase monotonically, followed by a phase in which it decreases monotonically. However, the loss can be highly non-monotone in the general case. We illustrate this in Fig. 3b, where we show the loss progression for FA* for the case where $m = n = 100$ and $r = 99$. In our simulations, we observe such highly non-monotonic behaviour when r is close to n. We observe a similar highly non-monotone behavior of loss for FA as well (see Fig. 5).

Therefore, we need a more careful analysis for the general case. From the FA* update equations (4), we get $\frac{d\ Z^T Z}{dt} = 0$. That is, $Z^T Z$ does not change with time. For this discussion, let us assume Z is initialized such that $Z(0)^T Z(0) = I$, which implies $Z^T Z = I$ throughout (the proof in the appendix holds for any full column rank Z).

Also, let R denote the residual matrix $(Y-\hat{Y})C^T$, A denote the alignment matrix $CW^T + WC^T$, and ℓ denote the loss $\left\|R\right\|_F^2$. R_i denotes the ith row of R (viewed as a column vector). Using basic matrix calculus, we get

$$\frac{d\ell}{dt} = -Tr(RAR^T) = -\sum_{i=1}^n R_i^T A R_i, \qquad (5)$$

$$\frac{dA}{dt} = 2R^T R. \qquad (6)$$

Equation 5 says that if A is positive semi-definite (PSD), then the loss ℓ decreases with time. Equation 6 says that A becomes more PSD with time, that is, $x^T A x$ never decreases for any fixed x (see Fig. 4a). This is the sense in which alignment between C and W increases monotonically. However, unlike Example 1, this is not sufficient to claim that loss starts decreasing monotonically after some time. This is because A
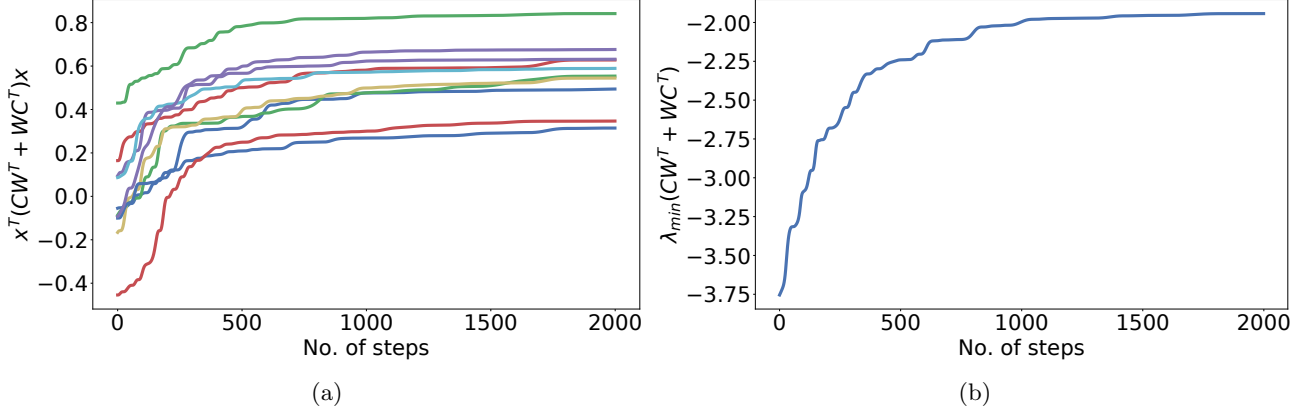
Figure 4: (a) $x^T(CW^T + WC^T)x$ vs time for 10 randomly chosen x. $x^T(CW^T + WC^T)x$ is monotonically increasing for all x. (b) Minimum eigenvalue of $CW^T + WC^T$ is monotonically increasing but can stay negative.

may never become PSD as there can be some x for which $\frac{d\, x^T A x}{dt}$ remains 0 after some time. We illustrate this in Fig. 4b where we show an instance where the minimum eigenvalue of A is monotonically increasing, but stays negative. That is, A does not become PSD. However, ℓ still converges to zero (Fig. 3b shows the corresponding loss progression).

To get past this hurdle, we need to understand the directions x for which $\frac{d\, x^T A x}{dt} > 0$. Observe that when $\frac{d\ell}{dt} > 0$, there is some row R_j of R such that $R_j^T A R_j < 0$. And $x^T A x$ increases sufficiently for all x for which $R_j^T x$ is large enough. In other words, when the loss increases, C and W become better aligned with respect to the direction which led to increase in loss (R_j), and all directions close to it. And such an R_j — with respect to which C and W are not aligned, satisfying $R_j^T A R_j < 0$ — must exist whenever the loss increases. Therefore, the loss can not increase indefinitely. Using this idea, we bound the total possible increase in loss, $\int_0^T \frac{d\ell}{dt} \mathbb{1}\left[\frac{d\ell}{dt} \geq 0\right]$ dt, for all T. Here, $\mathbb{1}[\cdot]$ denotes the indicator function which is equal to 1 if the condition inside the bracket is true, and 0 otherwise.

Using a similar argument, we can bound the total time for which the loss is large and is either increasing or decreasing very slowly. That is, we bound $\int_0^T \mathbb{1}\left[\ell \geq \epsilon \text{ and } \frac{d\ell}{dt} \geq -\delta\right] dt$ for all $\epsilon > 0, \delta > 0, T \geq 0$. At any other time, if the loss is large, it has to decrease sharply.

Therefore, the loss cannot increase too much, and cannot be in a slowly decreasing phase for too long. Using this, we show a bound on the time by which loss goes below ϵ (first part of Theorem 1). Building on these ideas, we can also show that the loss converges to 0 eventually. We refer the readers to Appendix A for more details.

In summary, here is the crux of the argument: whenever a bad event happens (increase in loss or slow decrease in loss), the alignment between C and W increases with respect to the direction which caused the bad event (R_j), and all directions close to it. Such a bad event can not happen when C and W are sufficiently aligned with respect to all rows of R. Therefore a bad event can not happen many times. This identifies the directions with respect to which alignment increases, and how this phenomenon facilitates convergence.

Implications for FA. The only difference between FA and FA* is that we set W optimally in the FA* update whereas we take the gradient step for W in the FA update. As we discussed in Example 1, the dynamics for FA and FA* can generally be very different. However, if we initialize FA with the optimal W (for the given Z) at $t = 0$, we observe that its loss progression is similar to FA* (see Fig. 5). The similarity is even more apparent if we initialize W optimally and choose a larger learning rate for W (compared to Z) in which case W continues to be close to optimal throughout the dynamics. For reference, we also include a plot for FA with randomly initialized W in Appendix F.

More generally, we believe that the ideas behind understanding alignment for FA* may also be helpful for FA. To see this, observe that for FA,

$$\frac{dA}{dt} = C(Y - \hat{Y})^T Z + Z^T(Y - \hat{Y})C^T,$$

$$\frac{d^2 A}{dt^2} = 2R^T R$$
$$- Z^T(Y - \hat{Y})C^T W C^T - C W C^T(Y - \hat{Y})^T Z$$
$$- Z^T Z Z^T(Y - \hat{Y})C^T - C(Y - \hat{Y})^T Z Z^T Z.$$

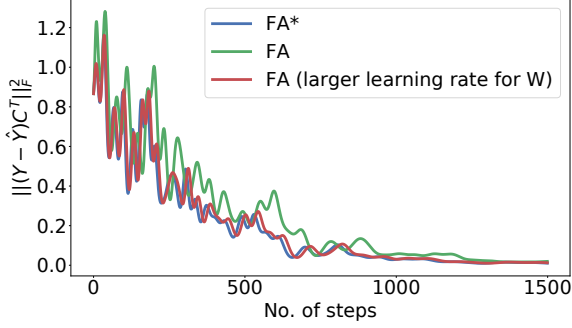Here, $A = CW^T + WC^T$ and $R = (Y - \hat{Y})C^T$. If W

Figure 5: FA and FA* loss progression when the initial value of W is chosen optimally for FA.

is close to optimal, then $Z^T(Y - \hat{Y}) \approx 0$, $dA/dt \approx 0$ and $d^2A/dt^2 \approx 2R^T R$. Recall from Equation 6 that $dA/dt = 2R^T R$ for FA*. This $R^T R$ term is the main reason for alignment. Thus, when $Z^T(Y - \hat{Y}) \approx 0$, one can hope to argue that A becomes more PSD with time and the ideas behind the analysis of FA* may be helpful for analysing FA. In general, if one can understand the progression of $Z^T(Y - \hat{Y})$, combining it with our insights about FA* mays yield a rigorous understanding of FA alignment dynamics.

4 UNDERSTANDING THE STATIONARY POINTS

In the previous section, we saw that FA* converges to a stationary point satisfying $(Y - \hat{Y})C^T = 0$, where $\hat{Y}_{n \times m} = Z_{n \times r} W_{r \times m}$ and $W = (Z^T Z)^{-1} Z^T Y$. In this section, we study the stationary points of feedback alignment and compare them to those of gradient flow (2). From prior work (Bah et al., 2019, Theorem 39, part (b)), we know that gradient flow starting from a random initialization converges to the optimal solution almost surely. We investigate when the solution found by feedback alignment is optimal, and how different the representations found using feedback alignment and gradient flow can be.

Note that the stationary point equations are same for FA and FA*, and are given by

$$
\begin{aligned}
(Y - \hat{Y})C^T &= 0 \\
Z^T(Y - \hat{Y}) &= 0 \\
\hat{Y} &= ZW.
\end{aligned}
\tag{7}
$$

So the results of this section apply to both versions of feedback alignment.

Characterization of stationary points. In the next lemma, we characterize the solution $\hat{Y}$ at stationary points when C is chosen randomly.

Lemma 1. *Suppose C is chosen randomly with entries drawn i.i.d. from $\mathcal{N}(0, 1)$ and the stationary point*

equations for feedback alignment (7) are satisfied. Let $A_{n \times r} = YC^T$ and $B_{r \times m} = \arg\min_B \|AB - Y\|_F^2$. Then $\hat{Y} = AB$ almost surely.

The proof of Lemma 1 can be found in Appendix B. To understand Lemma 1, we write $Y = \sum_{i=1}^n \sigma_i u_i v_i^T$ where σ_i is the ith singular value of Y, and u_i and v_i are the corresponding left and right singular vectors respectively. Then the jth column of A,

$$
A^{(j)} = \sum_{i=1}^n \sigma_i u_i R_{ij},
$$

where $R_{ij} = \langle v_i, C^{(j)} \rangle$ is a $\mathcal{N}(0, 1)$ random variable. That is, $A^{(j)}$ is a random linear combination of singular vectors of Y, scaled by its singular values. Lemma 1 says that feedback alignment finds the solution $\hat{Y}$ that corresponds to the best approximation of Y (in Frobenius norm) in the space spanned by $A^{(j)}$s. On the other hand, gradient descent finds the solution that corresponds to the best approximation of Y in the space spanned by top-r singular vectors of Y, which is also the optimal solution (see e.g., (Blum et al., 2020)).

Optimality of solution. Next, we show that when $r \geq rank(Y)$, feedback alignment stationary points correspond to the optimal solution (see Fig. 2a for an illustration).

Theorem 2. *Suppose C is chosen randomly with entries drawn i.i.d. from $\mathcal{N}(0, 1)$, the stationary point equations for feedback alignment (7) are satisfied, and $r \geq rank(Y)$. Then $ZW = \hat{Y} = Y$ almost surely, which also minimizes $\|ZW - Y\|_F^2$.*

The proof of Theorem 2 follows directly from Lemma 1 and can be found in Appendix B. Columns of A correspond to random linear combinations of singular vectors of Y, scaled by its singular values. We have at least $rank(Y)$ such columns. Therefore, the columns of A span the singular vectors of Y (corresponding to non-zero singular values) almost surely. Theorem 2 follows since $\hat{Y}$ is the best approximation of Y in the column span of A, which is equal to Y almost surely.

Also, note that this result does not hold for arbitrary C. For instance, suppose Y is a rank-1 matrix and $r = 1$. Let Z be any arbitrary full column-rank matrix, and $W = (Z^T Z)^{-1} Z^T Y$. In this case, $Y - \hat{Y}$ has rank at most 2. If we choose $C_{1 \times m}$ such that its only row is orthogonal to the row space of $Y - \hat{Y}$, then $(Y - \hat{Y})C^T = 0$ and the stationary point equations (7) are satisfied. However, $\hat{Y}$ may not be equal to Y. This motivates the random choice of C.

Next, we show that the feedback alignment solution can be far from optimal when r is much smaller than

$rank(Y)$ (see Fig. 2b for an illustration).

Theorem 3. *Suppose*
(i) C has entries drawn i.i.d. from $\mathcal{N}(0,1)$,
(ii) the stationary point equations for feedback alignment (7) are satisfied,
(iii) the singular values of Y satisfy

$$\sigma_i = \begin{cases} \frac{1}{\sqrt{2r}}, & for \ \ i \leq r \\ \frac{1}{\sqrt{2(n-r)}}, & for \ \ r+1 \leq i \leq n \end{cases}$$

(iv) $c_1 \leq r \leq c_2 n$ for some absolute constants c_1, c_2,

then the error $\|ZW - Y\|_F^2 \geq 0.74$ with probability at least 0.99 over the choice of C.

On the other hand, gradient flow (2) starting from randomly initialized Z and W (with i.i.d. $\mathcal{N}(0,1)$ entries) converges to the optimum solution with $\|ZW - Y\|_F^2 = 0.5$ almost surely.

Note that in the setting of Theorem 3, even a naive solution that sets Z and W to 0 obtains an error of 1, while feedback alignment and gradient descent errors are 0.74 and 0.5 respectively. In this sense, the feedback alignment solution is far from optimal. The proof of Theorem 3 can be found in Appendix C.

To understand Theorem 3, it is instructive to consider the case when $r = 1$. Let the singular values of Y satisfy

$$\sigma_i = \begin{cases} \frac{1}{\sqrt{2}}, & for \ \ i = 1 \\ \frac{1}{\sqrt{2(n-1)}}, & for \ \ 2 \leq i \leq n. \end{cases}$$

From Lemma 1, we know that $\hat{Y} = A_{n \times 1} B_{1 \times m}$ almost surely. Here A is a column vector satisfying

$$A = \frac{1}{\sqrt{2}} u_1 R_1 + \frac{1}{\sqrt{2(n-1)}} \sum_{i=2}^{n} u_i R_i,$$

where u_i are the left singular vectors of Y and R_i are drawn i.i.d. from $\mathcal{N}(0,1)$. On the other hand, the optimum solution corresponds to $A = u_1$. So while the optimum A aligns with the top singular vector, the A corresponding to feedback alignment has a significant component in the orthogonal subspace. This causes the feedback alignment solution to be far from optimal.

Comparison of representations. In the previous result, we saw that the error achieved by the feedback alignment solution can be much higher than the gradient flow solution. Next, we demonstrate that even when the two errors are approximately equal, the representations recovered by the two algorithms can be almost orthogonal, again in the rank-deficient setting.

Theorem 4. *Suppose*
(i) C has entries drawn i.i.d. from $\mathcal{N}(0,1)$,
(ii) the stationary point equations for feedback alignment (7) are satisfied,
(iii) the singular values of Y satisfy $\sigma_1 = 1$ and $\sigma_i = \epsilon$ for $i > 1$, where $0 < \epsilon < 1$,
(iv) $r = 1$ (rank 1 approximation) and $n \geq c$ for some absolute constant c.

Let Z_{FA} and W_{FA} denote the Z and W satisfying the above conditions respectively, and Z_{GD} and W_{GD} represent the factors found by gradient flow (2) starting from randomly initialized Z and W (with i.i.d. $\mathcal{N}(0,1)$ entries). Then

$$\|Z_{FA}W_{FA} - Y\|_F^2 \leq \|Z_{GD}W_{GD} - Y\|_F^2 \left(1 + \frac{2}{\epsilon^2 n}\right)$$

and

$$\left| \left\langle \frac{Z_{FA}}{\|Z_{FA}\|_2}, \frac{Z_{GD}}{\|Z_{GD}\|_2} \right\rangle \right| \leq \frac{4}{\epsilon\sqrt{n}}$$

with probability at least 0.99 over the choice of C and random initialization of gradient flow.

The proof of Theorem 4 can be found in Appendix D.

To understand Theorem 4, let us set $\epsilon = 0.5$. We get that the error of feedback alignment solution is at most $1 + O\left(\frac{1}{n}\right)$ times that of the gradient flow solution, while Z_{FA} and Z_{GD} are almost orthogonal, with normalized absolute inner product $O\left(\frac{1}{\sqrt{n}}\right)$.

From Lemma 1, we know that $\hat{Y} = A_{n \times 1} B_{1 \times m}$ almost surely. Here A is a column vector satisfying

$$A = u_1 R_1 + 0.5 \sum_{i=2}^{n} u_i R_i,$$

where u_i are the left singular vectors of Y and R_i are drawn i.i.d. from $\mathcal{N}(0,1)$. Since Z_{FA} and A are column vectors and $Z_{FA}W_{FA} = AB$ almost surely, we get

$$\frac{Z_{FA}}{\|Z_{FA}\|_2} = \frac{u_1 R_1 + 0.5 \sum_{i=2}^{n} u_i R_i}{\sqrt{R_1^2 + 0.25 \sum_{i=2}^{n} R_i^2}}$$

almost surely. Since gradient flow converges to the optimum solution almost surely, we know

$$\frac{Z_{GD}}{\|Z_{GD}\|_2} = u_1$$

almost surely. Using concentration of a χ-squared random variable, we get that the normalized absolute inner product between Z_{GD} and Z_{FA} is $O\left(\frac{1}{\sqrt{n}}\right)$ with high probability.

The error of gradient flow solution $\|Z_{GD}W_{GD} - Y\|_F^2$ is equal to the optimum error which is $\sum_{i=2}^{n} \sigma_i^2 =$

$0.25(n-1)$. It is not hard to see that the error of feedback alignment solution is at most $\|Y\|_F^2$ which is equal to $1 + 0.25(n-1)$. From here, we get that the error of feedback alignment solution is at most $1 + O\left(\frac{1}{n}\right)$ times that of the gradient flow solution.

Therefore, the errors of feedback alignment solution and the gradient flow solution can be approximately equal, while the representations they find are almost orthogonal. We note that for low-rank matrix factorization, this phenomenon happens only when the optimum error is large. For instance, when the optimum error is 0, we are in the regime where $r \geq rank(Y)$. In this case, the column space of Z_{FA} and Z_{GD} are equal to the column space of Y almost surely. It would be interesting to understand to what extent the representations found by gradient flow and feedback alignment are different for other problems such as for learning neural networks. And are there problems for which the representations are significantly different even with small optimum error?

5 LINEAR NEURAL NETWORKS

As a direct corollary, all our results for matrix factorization also hold for training two-layer linear neural networks, assuming the training input is isotropic and the output is a linear function of the input. Specifically, let $O = XY$, where the rows X_i and O_i represent the ith training input and output respectively, and let $X^T X = I$. We want to find Z and W that minimize the training error

$$\|XZW - O\|_F^2, \tag{8}$$

which is equal to the matrix factorization error $\|ZW - Y\|_F^2$ (Equation 1). We use $X^T X = I$ here, and in the update equations below. The gradient flow (GD) update for this problem is given by

$$\frac{dZ}{dt} = X^T(O - XZW)W^T = (Y - \hat{Y})W^T$$
$$\frac{dW}{dt} = Z^T X^T(O - XZW) = Z^T(Y - \hat{Y}). \tag{9}$$

where $\hat{Y} = ZW$. The feedback alignment (FA) update is given by

$$\frac{dZ}{dt} = X^T(O - XZW)C^T = (Y - \hat{Y})C^T$$
$$\frac{dW}{dt} = Z^T X^T(O - XZW) = Z^T(Y - \hat{Y}). \tag{10}$$

The update for feedback alignment with optimal W (FA*) is given by

$$\frac{dZ}{dt} = X^T(O - XZW)C^T = (Y - \hat{Y})C^T$$
$$W = (Z^T X^T X Z)^{-1} Z^T X^T O = (Z^T Z)^{-1} Z^T Y \tag{11}$$

As the GD, FA and FA* updates, and the error term are same as the corresponding updates and the error term for matrix factorization (Equations 2, 3, 4), all our results also hold in this case.

6 CONCLUSION

We investigate how and when feedback alignment works, focusing on the problem of low-rank matrix factorization. For the "how" question, we studied the dynamics of alignment between forward and backward weights, and its implications for convergence. For the "when" question, we showed that feedback alignment converges to the optimal solution when the factorization has rank $r \geq rank(Y)$, but it can be far from optimal in the rank-deficient case where $r < rank(Y)$. To the best of our knowledge, this is the first rigorous separation result between feedback alignment and gradient descent. We also demonstrate that the representations learned by feedback alignment and gradient descent can be very different, even when their errors are approximately equal.

There are many interesting directions for future research. A natural next step is to extend our understanding of alignment dynamics to the problem of learning non-linear neural networks. Song et al. (2021) show that alignment may not happen in highly overparameterized neural networks. But it is a robust phenomenon in the parameter regimes typically encountered in practice, and therefore important to understand. It would also be interesting to understand the implicit regularization properties of feedback alignment and compare them to gradient descent by considering problems such as matrix sensing in the overparameterized regime (Gunasekar et al., 2017; Li et al., 2018). From the point of view of the theory of optimization, a fundamental question is whether feedback alignment is part of a larger family of algorithms (e.g., that replace parts of the gradient with random values) and whether it might be applicable to problems even without layered structure. More generally, building the mathematical foundations of biologically plausible learning is a fruitful direction that can reveal surprising algorithms while advancing our understanding of the brain.

Acknowledgements

References

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):1–10, 2016.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

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

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

Arild Nøkland. Direct feedback alignment provides learning in deep neural networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 1045–1053, 2016.

Julien Launay, Iacopo Poli, François Boniface, and Florent Krzakala. Direct feedback alignment scales to modern deep learning tasks and architectures. *Advances in Neural Information Processing Systems*, 33, 2020.

Sergey Bartunov, Adam Santoro, Blake A Richards, Luke Marris, Geoffrey E Hinton, and Timothy P Lillicrap. Assessing the scalability of biologically-motivated deep learning algorithms and architectures. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 9390–9400, 2018.

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

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Simon S Du, Wei Hu, and Jason D Lee. Algorithmic regularization in learning deep homogeneous models:

Layers are automatically balanced. *arXiv preprint arXiv:1806.00900*, 2018.

Hossein Valavi, Sulin Liu, and Peter Ramadge. Revisiting the landscape of matrix factorization. In *International Conference on Artificial Intelligence and Statistics*, pages 1629–1638. PMLR, 2020.

Tian Ye and Simon S Du. Global convergence of gradient descent for asymmetric low-rank matrix factorization. *arXiv preprint arXiv:2106.14289*, 2021.

Bubacarr Bah, Holger Rauhut, Ulrich Terstiege, and Michael Westdickenberg. Learning deep linear neural networks: Riemannian gradient flows and convergence to global minimizers. *arXiv preprint arXiv:1910.05505*, 2019.

Pierre Baldi, Peter Sadowski, and Zhiqin Lu. Learning in the machine: Random backpropagation and the deep learning channel. *Artificial intelligence*, 260:1–35, 2018.

Ganlin Song, Ruitu Xu, and John Lafferty. Convergence and alignment of gradient descentwith random back propagation weights. *arXiv preprint arXiv:2106.06044*, 2021.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572*, 2018.

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*, pages 8925–8935. PMLR, 2021.

Timothy P Lillicrap, Adam Santoro, Luke Marris, Colin J Akerman, and Geoffrey Hinton. Backpropagation and the brain. *Nature Reviews Neuroscience*, 21(6):335–346, 2020.

Mohamed Akrout, Collin Wilson, Peter C Humphreys, Timothy Lillicrap, and Douglas Tweed. Deep learning without weight transport. *arXiv preprint arXiv:1904.05391*, 2019.

Yoshua Bengio. How auto-encoders could provide credit assignment in deep networks via target propagation. *arXiv preprint arXiv:1407.7906*, 2014.

Dong-Hyun Lee, Saizheng Zhang, Asja Fischer, and Yoshua Bengio. Difference target propagation. In *Joint european conference on machine learning and knowledge discovery in databases*, pages 498–515. Springer, 2015.

Daniel Kunin, Jonathan Bloom, Aleksandrina Goeva, and Cotton Seed. Loss landscapes of regularized linear autoencoders. In *International Conference on Machine Learning*, pages 3560–3569. PMLR, 2019.

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*, pages 5511–5521. PMLR, 2020.

Xiaohui Xie and H Sebastian Seung. Equivalence of backpropagation and contrastive hebbian learning in a layered network. *Neural computation*, 15(2):441–454, 2003.

Yoshua Bengio, Thomas Mesnard, Asja Fischer, Saizheng Zhang, and Yuhuai Wu. Stdp-compatible approximation of backpropagation in an energy-based model. *Neural computation*, 29(3):555–577, 2017.

Benjamin Scellier and Yoshua Bengio. Equilibrium propagation: Bridging the gap between energy-based models and backpropagation. *Frontiers in computational neuroscience*, 11:24, 2017.

James CR Whittington and Rafal Bogacz. An approximation of the error backpropagation algorithm in a predictive coding network with local hebbian synaptic plasticity. *Neural computation*, 29(5):1229–1262, 2017.

Jordan Guerguiev, Timothy P Lillicrap, and Blake A Richards. Towards deep learning with segregated dendrites. *Elife*, 6:e22901, 2017.

João Sacramento, Rui Ponte Costa, Yoshua Bengio, and Walter Senn. Dendritic cortical microcircuits approximate the backpropagation algorithm. *arXiv preprint arXiv:1810.11393*, 2018.

Christos H Papadimitriou, Santosh S Vempala, Daniel Mitropolsky, Michael Collins, and Wolfgang Maass. Brain computation by assemblies of neurons. *Proceedings of the National Academy of Sciences*, 117(25):14464–14472, 2020.

Max Dabagia, Christos H. Papadimitriou, and Santosh S. Vempala. Assemblies of neurons can learn to classify well-separated distributions, 2021.

Avrim Blum, John Hopcroft, and Ravindran Kannan. *Foundations of data science*. Cambridge University Press, 2020.

Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. *Advances in Neural Information Processing Systems*, 30, 2017.

Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*, pages 2–47. PMLR, 2018.

Martin Wainwright. Basic tail and concentration bounds. *URl: https://www. stat. berkeley. edu/.../Chap2_TailBounds_Jan22_2015. pdf (visited on 12/31/2017)*, 2015.

Supplementary Material:
How and When Random Feedback Works:
A Case Study of Low-Rank Matrix Factorization

A PROOF OF THEOREM 1

Let $Y_{n \times m}$ be the matrix we want to factorize and $\hat{Y}_{n \times k} = Z_{n \times r} W_{r \times m}$. Let $C_{r \times m}$ be the feedback matrix. Feedback alignment (FA*) updates Z and W as follows:

$$\frac{dZ}{dt} = (Y - \hat{Y})C^T$$
$$W = (Z^T Z)^{-1} Z^T Y.$$

Notation. We will use $M(t)$ to denote matrix M at time t. However, we will not show time t when it is clear from context. For a symmetric matrix M, we use $\lambda_i(M)$ to denote the ith largest eigenvalue of M. For any vector v, we use $\|v\|$ to denote the ℓ_2 norm of v. For any matrix M, we use M^i to denote its ith row.

We use $\mathbb{1}[.]$ to denote the indicator function which is equal to 1 if the condition inside the square brackets is true and 0 otherwise.

We use A to denote the alignment matrix $\left(\left(Z^T Z \right)^{-1} C W^T + W C^T \left(Z^T Z \right)^{-1} \right)$ and R to denote the residual $(Y - \hat{Y})C^T$.

For a non-zero vector $x(t)$, we use $x_{\leq k}(t)$ to denote vector $x(t)$ if $\frac{x(t)^T A(t) x(t)}{\|x(t)\|^2} \leq k$, and zero vector otherwise. Similarly, we use $x_{>k}(t)$ to denote vector $x(t)$ if $\frac{x(t)^T A(t) x(t)}{\|x(t)\|^2} > k$, and zero vector otherwise. We define $x_{\leq k}(t)$ and $x_{>k}(t)$ to be equal to $x(t)$ when $x(t)$ is a zero vector. For a matrix $M(t)$ with i^{th} row $M^i(t)$, $M_{\leq k}(t)$ denotes the matrix whose i^{th} row equals $M_{\leq k}^i(t)$ for all i. Similarly, we define $M_{>k}(t)$ to be the matrix whose i^{th} row equals $M_{>k}^i(t)$ for all i. Note that we can write $x(t) = x_{\leq k}(t) + x_{>k}(t)$ and $M(t) = M_{\leq k}(t) + M_{>k}(t)$.

We will use the loss function $\ell(t) = \left\| (Y - \hat{Y}(t))C^T \left(Z(t)^T Z(t) \right)^{-1/2} \right\|_F^2$.

Theorem 1. *Let $Z(0)$ be full column rank. For any $\epsilon > 0$ and*

$$T \geq \frac{24}{\epsilon} \left(\frac{\sigma_1(Y) \sigma_1(C) \sigma_1(Z(0))^6 \sqrt{r \ min(m,n)}}{\sigma_r(Z(0))^5} \right),$$

FA dynamics (4) satisfy*

$$\min_{t \leq T} \left\| (Y - \hat{Y}(t))C^T \right\|_F^2 \leq \epsilon.$$

Moreover,

$$\lim_{t \to \infty} \left\| (Y - \hat{Y}(t))C^T \right\|_F^2 = 0.$$

Note on the bound on T. The bound on T depends on the condition number of $Z(0)$ and the top singular values of Y, C and $Z(0)$. To understand this bound, suppose we set $\epsilon = \epsilon_1 \|Y\|_F^2 \|C\|_F^2$ (for some $\epsilon_1 > 0$), so

that ϵ has same scale as $\left\| (Y - \hat{Y})C^T \right\|_F^2$. Then the bound on T is given by

$$\frac{24}{\epsilon_1} \left(\frac{\sigma_1(Y)\,\sigma_1(C)\,\sigma_1(Z(0))^6\,\sqrt{r\,min(m,n)}}{\|Y\|_F^2\,\|C\|_F^2\,\sigma_r(Z(0))^5} \right).$$

This bound decreases if we scale up Y and C and scale down $Z(0)$. This is because $\frac{dZ}{dt} = (Y - \hat{Y})C^T$. So the relative magnitude of update to Z increases if we scale up Y and C and scale down $Z(0)$. The bound obtained would not be dependent on the scales of $Z(0)$, C, and Y, if we choose a scale independent update given by $\frac{dZ}{dt} = \frac{(Y - \hat{Y})C^T \|Z\|_F}{\|Y\|_F \|C\|_F}$.

A.1 Proof Overview

From Fact 3, we know that $Z^T Z$ does not change with time. While our formal proof holds for arbitrarily initialized Z (with full column rank), in this proof sketch, we will assume that Z is initialized such that $Z^T Z = I$. In Lemma 4, we show that

$$\frac{d\ell}{dt} = -Tr(RAR^T)$$

$$= -\sum_{i=1}^{n} R^{i^T} A R^i$$

where R and A are residual and alignment matrices respectively, as defined above, and R^i is the ith row (viewed as a column vector) of R. This implies that $\frac{d\ell}{dt} \leq 0$ if A is PSD. In Lemma 5, we show that

$$\frac{dA}{dt} = 2(Z^T Z)^{-1} R^T R (Z^T Z)^{-1}$$

$$= 2R^T R.$$

This implies that $x^T A x$ never decreases with time for all x. In this sense, A becomes more PSD with time. However, this is not sufficient to claim that A will become PSD eventually as there can exist x for which $\frac{dx^T A x}{dt} = 0$ at all times. Therefore, a more careful analysis of the directions in which A becomes PSD (x such that $x^T A x \geq 0$) is needed.

Whenever $\frac{d\ell}{dt}$ is positive, there is some row R^i of R for which $R^{i^T} A R^i$ is negative. Also, note that $\frac{d\,x^T A x}{dt} = 2\|Rx\|^2 \geq 2(R^{i^T}x)^2 > 0$, for x such that $R_i^T x \neq 0$. So whenever a direction R^i causes the loss to increase sufficiently, $x^T A x$ also increases sufficiently for all x close to R^i. And when $x^T A x > 0$ for all x, the loss can not increase anymore. That is, whenever some direction causes the loss to increase, A becomes "more PSD" for all directions close to this direction, and when A is PSD for all directions, the loss can not increase anymore. Using this idea, we bound the total increase in loss possible $\left(\int_0^T \frac{d\ell}{dt} \mathbb{1}\left[\frac{d\ell}{dt} \geq 0\right]\,dt \right)$ in Lemma 8. Using a similar idea, in lemma 9, we upper bound the total time for which the following holds: the loss is large and the loss is either increasing or decreasing slowly. At any other time, if the loss is large, it has to decrease sharply. Combining these two lemmas, in Lemma 10, we show a bound on time by which the loss goes below ϵ at least once. In Lemma 11, we optimize the bound proved in Lemma 10. In Lemma 12, we translate the guarantee on $\ell(t)$ to the desired guarantee on $\left\| (Y - \hat{Y})C^T \right\|_F^2$. The proves the first part of the theorem.

The results in Lemma 8 and 9 crucially rely on Lemma 7 which gives an upper bound on

$$\int_0^T \|R_{\leq k}(t)\|_F^2\,dt.$$

for all $k \geq 0$ and for all T. This lemma helps formalize the intuition discussed above. An upper bound on $\int_0^T \|R_{\leq 0}(t)\|_F^2\,dt$ lets us upper bound the total increase in loss (Lemma 9). An upper bound on $\int_0^T \|R_{\leq k}(t)\|_F^2\,dt$ for positive k lets us bound the total time for which the loss is large and is either increasing or decreasing slowly (Lemma 7).

Note that whenever $\|R_{\leq k}(t)x\|_2^2$ is large for any x, $x^T A(t)x$ increases by a large amount. Also, by definition, for any row $R_{\leq k}^i(T)$ (viewed as a column vector) of $R_{\leq k}(T)$, $R_{\leq k}^i(T)^T A(T)R_{\leq k}^i(T)$ can not be too large, that is, $R_{\leq k}^i(T)^T A(T)R_{\leq k}^i(T) \leq k\left\|R_{\leq k}^i(T)\right\|_2^2$. We also know that $R_{\leq k}^i(T)^T A(0)R_{\leq k}^i(T) \geq \lambda_r(A(0))\left\|R_{\leq k}^i(T)\right\|_2^2$. Therefore, for any row $R_{\leq k}^i(T)$,

$$
\begin{aligned}
\int_0^T \left\|R_{\leq k}(t)R_{\leq k}^i(T)\right\|_2^2 dt &\leq \int_0^T \left\|R(t)R_{\leq k}^i(T)\right\|_2^2 dt \\
&= \frac{1}{2}\int_0^T R_{\leq k}^i(T)^T \frac{dA(t)}{dt}R_{\leq k}^i(T)dt \\
&= \frac{1}{2}\left(R_{\leq k}^i(T)^T A(T)R_{\leq k}^i(T) - R_{\leq k}^i(T)^T A(0)R_{\leq k}^i(T)\right) \\
&\leq \frac{1}{2}(k - \lambda_r(A(0)))\left\|R_{\leq k}^i(T)\right\|_2^2
\end{aligned}
$$

In other words, this bounds the inner product of rows of $R_{\leq k}(T)$ with the rows of $R_{\leq k}(t)$ for $t \leq T$. This fact lets us upper bound the integral of sum of squared norms of these rows. We prove such a bound for general vectors in Lemma 6, and use it to prove Lemma 7.

We prove convergence of $\ell(t)$ to zero (second part of the theorem) in Lemma 13, where we use the following argument. In Lemma 10, we show that for all $\epsilon > 0$ and all $T \geq 0$, there exists $t \geq T$ such that $\ell(t) \leq \epsilon$. Now, if the loss doesn't converge to 0, then there must exist some $\epsilon_1 > 0$, such that for all $T \geq 0$, there exists some $t \geq T$ satisfying $\ell(t) > \epsilon_1$. Using these two arguments, we can generate an infinite increasing sequence $T_1, T_1', T_2, T_2', \cdots$ such that $\ell(T_i) \leq \epsilon_1/2$ and $\ell(T_i') > \epsilon_1$ for all i. Thus we can get infinitely many disjoint intervals $[T_i, T_i']$ on which the loss increases by at least $\epsilon_1/2$, implying that the total increase in loss is unbounded which contradicts Lemma 8, where we show $\int_0^T \frac{d\ell}{dt}\mathbb{1}\left[\frac{d\ell}{dt} \geq 0\right] dt$ is bounded for all T. Therefore, the loss $\ell(t)$ must converge to 0.

A.2 Proof

The following two facts will be useful for the proof.

Fact 2.
$$Z^T(Y - \hat{Y}) = 0.$$

Fact 3.
$$\frac{d(Z^T Z)}{dt} = 0.$$

Fact 2 follows since $\hat{Y} = ZW = Z(Z^T Z)^{-1}Z^T Y$. Fact 3 follows since $\frac{d(Z^T Z)}{dt} = Z^T(Y - \hat{Y})C^T + (Z^T(Y - \hat{Y})C^T)^T = 0$.

Fact 3 says that $Z^T Z$ does not change with time. While our result holds for arbitrarily initialized Z (with full column rank), it might be helpful for the reader to assume that Z is initialized such that $Z^T Z = I$.

Now, we evaluate the expression for $\frac{d\ell}{dt}$.

Lemma 4.
$$\frac{d\ell}{dt} = -Tr\left((Y - \hat{Y})C^T\left(\left(Z^T Z\right)^{-1}CW^T + WC^T\left(Z^T Z\right)^{-1}\right)C(Y - \hat{Y})^T\right).$$

Proof. We can write

$$\frac{d\ell}{dt} = Tr\left(\left(\frac{dZ}{dt}\right)^T\left(\frac{d\ell}{dZ}\right)\right). \tag{12}$$

where

$$\frac{d\ell}{dZ} = -2(Y - \hat{Y})C^T(Z^T Z)^{-1}CW^T - 2Z(Z^T Z)^{-1}C(Y - \hat{Y})^T(Y - \hat{Y})C^T(Z^T Z)^{-1}.$$

Here, we used $Z^T(Y - \hat{Y}) = 0$ (see Fact 2) to simplify the expression. Substituting this in equation 12, we get

$$\frac{d\ell}{dt} = Tr\left(\left(C(Y - \hat{Y})^T\right)\left(-2(Y - \hat{Y})C^T(Z^TZ)^{-1}CW^T - 2Z(Z^TZ)^{-1}C(Y - \hat{Y})^T(Y - \hat{Y})C^T(Z^TZ)^{-1}\right)\right).$$

Again using $Z^T(Y - \hat{Y}) = 0$, we get

$$\frac{d\ell}{dt} = -2Tr\left(C(Y - \hat{Y})^T(Y - \hat{Y})C^T\left((Z^TZ)^{-1}CW^T\right)\right)$$

Using the identities $Tr(MN) = Tr(NM)$ and $Tr(M^T) = Tr(M)$, we get

$$\frac{d\ell}{dt} = -2Tr\left((Y - \hat{Y})C^T\left((Z^TZ)^{-1}CW^T\right)C(Y - \hat{Y})^T\right)$$
$$= -Tr\left((Y - \hat{Y})C^T\left((Z^TZ)^{-1}CW^T + WC^T(Z^TZ)^{-1}\right)C(Y - \hat{Y})^T\right).$$

$\square$

Recall that the alignment matrix is $A = \left((Z^TZ)^{-1}CW^T + WC^T(Z^TZ)^{-1}\right)$ and the residual is $R = (Y - \hat{Y})C^T$. Next, we show how A changes with time.

Lemma 5.

$$\frac{dA}{dt} = 2(Z^TZ)^{-1}R^TR(Z^TZ)^{-1}$$

Proof. From Fact 3, we know that $\frac{d(Z^TZ)}{dt} = 0$. This implies

$$\frac{dA}{dt} = \left((Z^TZ)^{-1}C\frac{dW}{dt}^T + \frac{dW}{dt}C^T(Z^TZ)^{-1}\right) \quad (13)$$

Here, $W = (Z^TZ)^{-1}Z^TY$. Again, using $\frac{d(Z^TZ)}{dt} = 0$ and $\frac{dZ}{dt} = (Y - \hat{Y})C^T$, we get

$$\frac{dW}{dt} = (Z^TZ)^{-1}C(Y - \hat{Y})^TY.$$

From Fact 2, we know that $Z^T(Y - \hat{Y}) = 0$ which implies $(Y - \hat{Y})^T\hat{Y} = 0$. Using this, we get

$$\frac{dW}{dt} = (Z^TZ)^{-1}C(Y - \hat{Y})^T(Y - \hat{Y}).$$

Substituting this in Equation 13, we get

$$\frac{dA}{dt} = 2(Z^TZ)^{-1}C(Y - \hat{Y})^T(Y - \hat{Y})C^T(Z^TZ)^{-1}$$
$$= 2(Z^TZ)^{-1}R^TR(Z^TZ)^{-1}.$$

$\square$

The lemma below essentially says the following: suppose we observe n vectors $v_1(t), v_2(t) \cdots v_n(t)$ at each time t, and let the vectors observed at any time T have small inner product with all vectors observed before time T ($\sum_{i=1}^n \int_0^T \langle v_k(T), v_i(t)\rangle^2 dt$ is small), for all T. Then $\sum_{i=1}^n \int_0^T \|v_i(t)\|^2 dt$ can not be too large, for all T.

Lemma 6. *Let $v_i(t) : \mathbb{R} \to \mathbb{R}^r$ for $i \in \{1, 2 \cdots, n\}$ such that*

$$\sum_{i=1}^n \int_0^T \langle v_k(T), v_i(t)\rangle^2 dt \le c \|v_k(T)\|^2 \quad (14)$$

for all T, k and for some constant $c \ge 0$. Then

$$\sum_{i=1}^n \int_0^T \|v_i(t)\|^2 dt \le 2rc, \quad (15)$$

for all T (assuming $\|v_i(t)\|^2$ is integrable).

Proof. Let $L_T = \sum_{i=1}^{n} \int_0^T v_i(t) v_i(t)^T dt$. Sum of eigenvalues of L is given by

$$\sum_{i=1}^{r} \lambda_i(L_T) = \sum_{i=1}^{n} \int_0^T \|v_i(t)\|^2 \, dt. \tag{16}$$

Here, we used $\sum_{i=1}^{r} \lambda_i(L_T) = Tr(L_T)$. Using $\sum_{i=1}^{r} \lambda_i(L_T)^2 = Tr(L_T L_T^T)$, we get

$$
\begin{aligned}
\sum_{i=1}^{r} \lambda_i(L_T)^2 &= \sum_{i=1}^{n} \sum_{j=1}^{n} \int_0^T \int_0^T \langle v_i(t_1), v_j(t_2) \rangle^2 \, dt_1 \, dt_2 \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \int_0^T \int_0^{t_2} \langle v_i(t_1), v_j(t_2) \rangle^2 \, dt_1 \, dt_2 + \sum_{i=1}^{n} \sum_{j=1}^{n} \int_0^T \int_0^{t_1} \langle v_i(t_1), v_j(t_2) \rangle^2 \, dt_2 \, dt_1 \\
&= \sum_{j=1}^{n} \int_0^T \sum_{i=1}^{n} \int_0^{t_2} \langle v_i(t_1), v_j(t_2) \rangle^2 \, dt_1 \, dt_2 + \sum_{i=1}^{n} \int_0^T \sum_{j=1}^{n} \int_0^{t_1} \langle v_i(t_1), v_j(t_2) \rangle^2 \, dt_2 \, dt_1
\end{aligned}
$$

Using condition 14, we get

$$\sum_{i=1}^{r} \lambda_i(L_T)^2 \leq 2 \sum_{i=1}^{n} \int_0^T c \, \|v_i(t)\|^2 \, dt. \tag{17}$$

Using Cauchy–Schwarz inequality, we can write

$$\left(\sum_{i=1}^{r} \lambda_i(L_T) \right)^2 \leq r \sum_{i=1}^{r} \lambda_i(L_T)^2.$$

Substituting from Equation 16 and 17, we get

$$\left(\sum_{i=1}^{n} \int_0^T \|v_i(t)\|^2 \, dt \right)^2 \leq 2rc \sum_{i=1}^{n} \int_0^T \|v_i(t)\|^2 \, dt.$$

which implies

$$\sum_{i=1}^{n} \int_0^T \|v_i(t)\|^2 \, dt \leq 2rc.$$

In the last step, we used $\sum_{i=1}^{n} \int_0^T \|v_i(t)\|^2 \, dt \neq 0$. If it is equal to zero, then the lemma is trivially true. $\square$

Recall that A denotes the alignment matrix $\left((Z^T Z)^{-1} C W^T + W C^T (Z^T Z)^{-1} \right)$, and R denotes the residual matrix $(Y - \hat{Y}) C^T$.

Also recall the following notation. For a non-zero vector $x(t)$, we use $x_{\leq k}(t)$ to denote vector $x(t)$ if $\frac{x(t)^T A(t) x(t)}{\|x(t)\|^2} \leq k$, and zero vector otherwise. Similarly, we use $x_{>k}(t)$ to denote vector $x(t)$ if $\frac{x(t)^T A(t) x(t)}{\|x(t)\|^2} > k$, and zero vector otherwise. We define $x_{\leq k}(t)$ and $x_{>k}(t)$ to be equal to $x(t)$ when $x(t)$ is a zero vector. For a matrix $M(t)$ with i^{th} row $M^i(t)$, $M_{\leq k}(t)$ denotes the matrix whose i^{th} row equals $M_{\leq k}^i(t)$ for all i. Similarly, we define $M_{>k}(t)$ to be the matrix whose i^{th} row equals $M_{>k}^i(t)$ for all i. Note that we can write $x(t) = x_{\leq k}(t) + x_{>k}(t)$ and $M(t) = M_{\leq k}(t) + M_{>k}(t)$.

In the next lemma, we show an upper bound on $\int_0^T \|R_{\leq k}(t)\|_F^2 \, dt$. We will see in Lemma 8 and Lemma 9 that $\|R_{\leq k}(t)\|_F^2$ being large corresponds to certain undesirable events. For example, in Lemma 8, we will see that $\|R_{\leq 0}(t)\|_F^2$ being large corresponds to increase in loss $\ell(t)$. The next lemma will be helpful in bounding the total time for which such undesirable events can happen.

Lemma 7. *For all $k \geq \lambda_r(A(0))$ and for all T,*

$$\int_0^T \|R_{\leq k}(t)\|_F^2 \, dt \leq r \, (k - \lambda_r(A(0))) \, \lambda_1 \left(Z(0)^T Z(0) \right)^2.$$

Proof. Let $R_{\leq k}^i(T)$ be any row of $R_{\leq k}(T)$. By definition, we know that $R_{\leq k}^i(T)$ (viewed as a column vector) satisfies

$$(R_{\leq k}^i(T))^T A(T)(R_{\leq k}^i(T)) \leq k \|R_{\leq k}^i(T)\|^2 \tag{18}$$

for all i. Since $\lambda_r(A(0))$ is the minimum eigenvalue of $A(0)$, we also know that

$$(R_{\leq k}^i(T))^T A(0)(R_{\leq k}^i(T)) \geq \lambda_r(A(0)) \|R_{\leq k}^i(T)\|^2. \tag{19}$$

The last two equation imply

$$(R_{\leq k}^i(T))^T A(T)(R_{\leq k}^i(T)) - (R_{\leq k}^i(T))^T A(0)(R_{\leq k}^i(T)) \leq (k - \lambda_r(A(0))) \|R_{\leq k}^i(T)\|^2. \tag{20}$$

This implies

$$\int_0^T (R_{\leq k}^i(T))^T \left(\frac{d(A(t))}{dt}\right)(R_{\leq k}^i(T))dt \leq (k - \lambda_r(A(0)))\|R_{\leq k}^i(T)\|^2. \tag{21}$$

Substituting for $\frac{d(A(t))}{dt}$ from Lemma 5, we get

$$\int_0^T 2(R_{\leq k}^i(T))^T \left(\left(Z(t)^T Z(t)\right)^{-1} R(t)^T R(t) \left(Z(t)^T Z(t)\right)^{-1}\right)(R_{\leq k}^i(T)) \, dt \leq (k - \lambda_r(A(0)))\|R_{\leq k}^i(T)\|^2. \tag{22}$$

Note that for all x and k, $x^T R^T R x \geq x^T R_{\leq k}^T R_{\leq k} x$. Also using Fact 3, we know that $\left(Z(t)^T Z(t)\right)^{-1} = \left(Z(0)^T Z(0)\right)^{-1}$. This gives us

$$\int_0^T (R_{\leq k}^i(T))^T \left(\left(Z(0)^T Z(0)\right)^{-1} R_{\leq k}(t)^T R_{\leq k}(t) \left(Z(0)^T Z(0)\right)^{-1}\right)(R_{\leq k}^i(T)) \, dt \leq \frac{(k - \lambda_r(A(0))) \|R_{\leq k}^i(T)\|^2}{2}. \tag{23}$$

The expression in the above integral is integrable by Lemma 16.

Now, define $v_i(t) = \left(Z(0)^T Z(0)\right)^{-\frac{1}{2}} R_{\leq k}^i(t)$. The above equation implies

$$\int_0^T \sum_{j=1}^n \langle v_i(T), v_j(t)\rangle^2 \, dt \leq \frac{(k - \lambda_r(A(0))) \, v_i(T)^T (Z(0)^T Z(0)) v_i(T)}{2}. \tag{24}$$

We know $v_i(T)^T(Z(0)^T Z(0))v_i(T) \leq \lambda_1(Z(0)^T Z(0)) \|v_i(T)\|^2$. This gives us

$$\int_0^T \sum_{j=1}^n \langle v_i(T), v_j(t)\rangle^2 \, dt \leq \frac{(k - \lambda_r(A(0))) \, \lambda_1\left(Z(0)^T Z(0)\right) \|v_i(T)\|^2}{2}. \tag{25}$$

Using Lemma 6, we get

$$\int_0^T \sum_{j=1}^n \|v_j(t)\|^2 \, dt \leq r \, (k - \lambda_r(A(0))) \, \lambda_1\left(Z(0)^T Z(0)\right), \tag{26}$$

for all T ($\|v_j(t)\|^2$ is integrable by Lemma 16). This is equivalent to

$$\int_0^T \left\|R_{\leq k}(t)\left(Z(0)^T Z(0)\right)^{-\frac{1}{2}}\right\|_F^2 \, dt \leq r \, (k - \lambda_r(A(0))) \, \lambda_1\left(Z(0)^T Z(0)\right), \tag{27}$$

for all T. We know

$$\left\|R_{\leq k}(t)\left(Z(0)^T Z(0)\right)^{-\frac{1}{2}}\right\|_F^2 \geq \|R_{\leq k}(t)\|_F^2 \, \lambda_r\left(\left(Z(0)^T Z(0)\right)^{-\frac{1}{2}}\right)^2$$

$$= \frac{\|R_{\leq k}(t)\|_F^2}{\lambda_1\left((Z(0)^T Z(0))\right)} \tag{28}$$

Combining this with Equation 27, we get

$$\int_0^T \|R_{\leq k}(t)\|_F^2 \ dt \leq \ r \ (k - \lambda_r (A(0))) \ \lambda_1 \left(Z(0)^T Z(0)\right)^2 , \tag{29}$$

for all T ($\|R_{\leq k}(t)\|_F^2$ is integrable by Lemma 16).

This comples the proof. $\qquad\square$

Recall, we use $\mathbb{1}[.]$ to denote the indicator function which is equal to 1 if the condition inside the square brackets is true and 0 otherwise.

In the next lemma, we show an upper bound on total increase possible in loss $\ell(t)$.

Lemma 8. *For all T,*

$$\int_0^T \frac{d\ell(t)}{dt} \ \mathbb{1}\left[\frac{d\ell(t)}{dt} \geq 0\right] \ dt \leq r \ \lambda_r (A(0))^2 \ \lambda_1 \left(Z(0)^T Z(0)\right)^2 .$$

Proof. We do not explicitly show time t for variables in the proof below. Whenever time is not written, the corresponding variable is evaluated at some arbitary time t.

From Lemma 4 and Lemma 5, we know that

$$\frac{d\ell}{dt} = -Tr(RAR^T) \tag{30}$$

$$\frac{d(x^T A x)}{dt} = 2 \ x^T (Z^T Z)^{-1} R^T R (Z^T Z)^{-1} x \tag{31}$$

for all fixed x (which do not change with t). Here, R denotes the residual matrix $(Y - \hat{Y})C^T$ and A denotes the alignment matrix $\left(\left(Z^T Z\right)^{-1} CW^T + WC^T \left(Z^T Z\right)^{-1}\right)$.

Note that if $\lambda_r(A(0)) > 0$, then $\frac{d\ell}{dt} < 0$ for all t and the lemma is trivially true. In the proof below, we assume $\lambda_r(A(0)) \leq 0$.

By the definition of $R_{\leq k}$ and $R_{>k}$, we can write

$$\frac{d\ell}{dt} = -\left(Tr(R_{\leq 0} A R_{\leq 0}^T) + Tr(R_{>0} A R_{>0}^T)\right) . \tag{32}$$

By definition of $R_{>0}$, we know that $Tr(R_{>0} A R_{>0}^T) \geq 0$. This gives us

$$\begin{aligned}\frac{d\ell}{dt} &\leq -Tr(R_{\leq 0} A R_{\leq 0}^T) \\ &\leq -\lambda_r(A) \|R_{\leq 0}\|_F^2 .\end{aligned} \tag{33}$$

Here, $\lambda_r(A)$ is the minimum eigenvalue of $A(t)$. Since $\frac{d(x^T A x)}{dt} \geq 0$ for all fixed x, the minimum eigenvalue of A never decreases. That is, $\lambda_r(A(t)) \geq \lambda_r(A(0))$. This implies

$$\frac{d\ell}{dt} \leq -\lambda_r(A(0)) \|R_{\leq 0}\|_F^2 . \tag{34}$$

From Lemma 7, we know

$$\int_0^T \|R_{\leq 0}(t)\|_F^2 \ dt \leq - \ r \ \lambda_r (A(0)) \ \lambda_1 \left(Z(0)^T Z(0)\right)^2 , \tag{35}$$

for all T. Therefore we get

$$\begin{aligned}\int_0^T \frac{d\ell(t)}{dt} \ \mathbb{1}\left[\frac{d\ell(t)}{dt} \geq 0\right] \ dt &\leq \int_0^T -\lambda_r(A(0)) \|R_{\leq 0}(t)\|_F^2 \ \mathbb{1}\left[\frac{d\ell(t)}{dt} \geq 0\right] dt \\ &\leq -\lambda_r(A(0)) \int_0^T \|R_{\leq 0}(t)\|_F^2 \ dt \\ &\leq r \ \lambda_r (A(0))^2 \ \lambda_1 \left(Z(0)^T Z(0)\right)^2 ,\end{aligned} \tag{36}$$

for all T. Here, we used $\lambda_r(A(0)) \leq 0$. The expressions in the above integral are Lebesgue integrable as product of bounded integrable functions is integrable (over any finite interval). Here, $\frac{d\ell}{dt}$ is integrable due to continuity, $\mathbb{1}\left[\frac{d\ell(t)}{dt} \geq 0\right]$ is integrable by Lemma 15 and $\|R_{\leq 0}(t)\|_F^2$ is integrable by Lemma 16.

This completes the proof. $\qquad\square$

In the next lemma, we bound the total time for which loss $\|R\|_F$ is large, and ℓ is either increasing, or decreasing slowly. Since $\ell = \left\|R(Z^T Z)^{-1/2}\right\|_F^2$ and $Z^T Z$ doesn't change with time, this also gives a bound for total time for which loss ℓ is large, and ℓ is either increasing, or decreasing slowly.

Lemma 9. *For all $\delta > 0$, $\epsilon > 0$ and for all T,*

$$\int_0^T \mathbb{1}\left[\frac{d\ell(t)}{dt} > -\delta \text{ and } \|R(t)\|_F^2 > \epsilon\right] dt \leq \frac{r\left(\frac{2\delta}{\epsilon} - \lambda_r(A(0))\right)^2 \lambda_1\left(Z(0)^T Z(0)\right)^2}{\delta}.$$

Proof. We do not explicitly show time t for variables in the proof below. Whenever time is not written, the corresponding variable is evaluated at some arbitary time t.

From Lemma 4 and Lemma 5, we know that

$$\frac{d\ell}{dt} = -Tr(RAR^T), \tag{37}$$

$$\frac{d(x^T A x)}{dt} = 2 \; x^T (Z^T Z)^{-1} R^T R (Z^T Z)^{-1} x. \tag{38}$$

for all fixed x (which do not change with t). Here, R denotes the residual matrix $(Y - \hat{Y})C^T$ and A denotes the alignment matrix $\left((Z^T Z)^{-1} C W^T + W C^T (Z^T Z)^{-1}\right)$.

Note that if $\lambda_r(A(0)) > \frac{2\delta}{\epsilon}$, then $\frac{d\ell}{dt} < -2\delta$ for all t where $\|R(t)\|_F^2 \geq \epsilon$. In this case, the lemma is trivially true. In the proof below, we assume $\lambda_r(A(0)) \leq \frac{2\delta}{\epsilon}$.

Let $\mathbb{1}\left[\frac{d\ell(t)}{dt} > -\delta \text{ and } \|R(t)\|_F^2 > \epsilon\right] = 1$ at the current time t. Let $k = \frac{2\delta}{\epsilon}$. By the definition of $R_{\leq k}$ and $R_{>k}$, we can write

$$\frac{d\ell}{dt} = -\left(Tr(R_{\leq k} A R_{\leq k}^T) + Tr(R_{>k} A R_{>k}^T)\right). \tag{39}$$

Since $\frac{d\ell}{dt} > -\delta$, we get

$$\left(Tr(R_{\leq k} A R_{\leq k}^T) + Tr(R_{>k} A R_{>k}^T)\right) < \delta. \tag{40}$$

By definition of $R_{>k}$, we know that

$$Tr(R_{>k} A R_{>k}^T) \geq \|R_{>k}\|_F^2 \; k. \tag{41}$$

Since $\lambda_r(A)$ is the minimum eigenvalue of $A(t)$ and since the minimum eigenvalue of A does not decrease with time, we can write

$$\begin{aligned} Tr(R_{\leq k} A R_{\leq k}^T) &\geq \|R_{\leq k}\|_F^2 \; \lambda_r(A) \\ &\geq \|R_{\leq k}\|_F^2 \; \lambda_r(A(0)). \end{aligned} \tag{42}$$

Equation 40, 41 and 42 together imply

$$\|R_{\leq k}\|_F^2 \; \lambda_r(A(0)) + \|R_{>k}\|_F^2 \; k \leq \delta. \tag{43}$$

Since $\|R_{\leq k}\|_F^2 + \|R_{>k}\|_F^2 > \epsilon$ and $k = \frac{2\delta}{\epsilon}$, we get

$$\|R_{\leq k}\|_F^2 \; \lambda_r(A(0)) + \left(\epsilon - \|R_{\leq k}\|_F^2\right) \frac{2\delta}{\epsilon} \leq \delta. \tag{44}$$

Rearranging the terms, we get

$$\|R_{\leq k}\|_F^2 \geq \frac{\delta}{\frac{2\delta}{\epsilon} - \lambda_r(A(0))}. \tag{45}$$

This implies

$$\int_0^T \|R_{\leq k}(t)\|_F^2 \ dt \geq \int_0^T \|R_{\leq k}(t)\|_F^2 \ \mathbb{1}\left[\frac{d\ell(t)}{dt} > -\delta \text{ and } \|R(t)\|_F^2 > \epsilon\right] \ dt$$

$$\geq \frac{\delta}{\frac{2\delta}{\epsilon} - \lambda_r(A(0))} \int_0^T \mathbb{1}\left[\frac{d\ell(t)}{dt} > -\delta \text{ and } \|R(t)\|_F^2 > \epsilon\right] \ dt, \qquad (46)$$

for all T. The expressions in the above integral are Lebesgue integrable as product of bounded integrable functions is integrable (over any finite interval). Here, $\mathbb{1}\left[\frac{d\ell(t)}{dt} > -\delta\right]$ and $\mathbb{1}\left[\|R(t)\|_F^2 > \epsilon\right]$ are integrable by Lemma 15 and $\|R_{\leq k}(t)\|_F^2$ is integrable by Lemma 16.

From Lemma 7, we know

$$\int_0^T \|R_{\leq k}(t)\|_F^2 \ dt \leq \ r \ (k - \lambda_r(A(0))) \ \lambda_1\left(Z(0)^T Z(0)\right)^2$$

$$= \ r \ \left(\frac{2\delta}{\epsilon} - \lambda_r(A(0))\right) \ \lambda_1\left(Z(0)^T Z(0)\right)^2, \qquad (47)$$

for all T. From last two equations, we conclude

$$\int_0^T \mathbb{1}\left[\frac{d\ell(t)}{dt} > -\delta \text{ and } \|R(t)\|_F^2 > \epsilon\right] \ dt \leq \frac{r \left(\frac{2\delta}{\epsilon} - \lambda_r(A(0))\right)^2 \ \lambda_1\left(Z(0)^T Z(0)\right)^2}{\delta}, \qquad (48)$$

for all T. $\qquad\qquad\square$

Using above lemmas, in the next lemma, we show that the loss goes below ϵ at least once in every length T time interval for appropriately defined T.

Lemma 10. *For all $\epsilon > 0$, $\delta > 0$, $T_1 \geq 0$, and*

$$T \geq \frac{r \left(\frac{2\delta}{\epsilon \lambda_r(Z(0)^T Z(0))} - \lambda_r(A(0))\right)^2 \ \lambda_1\left(Z(0)^T Z(0)\right)^2}{\delta} + \frac{\ell(T_1) + \ r \ \lambda_r(A(0))^2 \ \lambda_1\left(Z(0)^T Z(0)\right)^2 - \epsilon}{\delta},$$

FA dynamics satisfy

$$\min_{T_1 \leq t \leq T_1 + T} \ell(t) \leq \epsilon.$$

Proof. Assume $\ell(t) > \epsilon$ for all $T_1 \leq t < T_1 + T$. We would show that $\ell(T_1 + T) \leq \epsilon$ in this case, which would imply the lemma.

We can write

$$\ell(T_1 + T) = \ell(T_1) + \int_{T_1}^{T_1+T} \frac{d\ell(t)}{dt} \ dt$$

$$= \ell(T_1) + \int_{T_1}^{T_1+T} \frac{d\ell(t)}{dt} \ \mathbb{1}\left[\frac{d\ell(t)}{dt} \leq -\delta\right] \ dt + \int_{T_1}^{T_1+T} \frac{d\ell(t)}{dt} \ \mathbb{1}\left[\frac{d\ell(t)}{dt} > -\delta\right] \ dt \qquad (49)$$

$$\leq \ell(T_1) + \int_{T_1}^{T_1+T} \frac{d\ell(t)}{dt} \ \mathbb{1}\left[\frac{d\ell(t)}{dt} \leq -\delta\right] \ dt + \int_{T_1}^{T_1+T} \frac{d\ell(t)}{dt} \ \mathbb{1}\left[\frac{d\ell(t)}{dt} \geq 0\right] \ dt.$$

The expressions in the above integral are Lebesgue integrable as product of bounded integrable functions is integrable (over any finite interval). Here, $\frac{d\ell}{dt}$ is integrable by continuity, $\mathbb{1}\left[\frac{d\ell(t)}{dt} > -\delta\right]$, $\mathbb{1}\left[\frac{d\ell(t)}{dt} \geq 0\right]$ and $\mathbb{1}\left[\frac{d\ell(t)}{dt} \leq -\delta\right]$ are integrable by Lemma 15.

We will now bound the two integral terms in the RHS. From Lemma 8, we know

$$\int_{T_1}^{T_1+T} \frac{d\ell(t)}{dt} \ \mathbb{1}\left[\frac{d\ell(t)}{dt} \geq 0\right] \ dt \leq \int_0^{T_1+T} \frac{d\ell(t)}{dt} \ \mathbb{1}\left[\frac{d\ell(t)}{dt} \geq 0\right] \ dt \qquad (50)$$

$$\leq r \ \lambda_r(A(0))^2 \ \lambda_1\left(Z(0)^T Z(0)\right)^2 \qquad (51)$$

We can write

$$\int_{T_1}^{T_1+T} \frac{d\ell(t)}{dt} \, \mathbb{1}\left[\frac{d\ell(t)}{dt} \leq -\delta\right] \, dt \leq -\delta \int_{T_1}^{T_1+T} \mathbb{1}\left[\frac{d\ell(t)}{dt} \leq -\delta\right] \, dt$$

$$= -\delta \left(T - \int_{T_1}^{T_1+T} \mathbb{1}\left[\frac{d\ell(t)}{dt} > -\delta\right] \, dt\right) \quad (52)$$

$$= -\delta \left(T - \int_{T_1}^{T_1+T} \mathbb{1}\left[\frac{d\ell(t)}{dt} > -\delta \text{ and } \ell(t) > \epsilon\right] \, dt\right)$$

where we used the assumption $\ell(t) > \epsilon$ for all $T_1 \leq t < T_1 + T$ in the last equation. Recall that $\ell(t) = \left\|R(t)(Z(t)^T Z(t))^{-\frac{1}{2}}\right\|_F^2 = \left\|R(t)(Z(0)^T Z(0))^{-\frac{1}{2}}\right\|_F^2$. Therefore, we can write

$$\epsilon < \left\|R(t)(Z(0)^T Z(0))^{-\frac{1}{2}}\right\|_F^2$$

$$< \|R(t)\|_F^2 \, \lambda_1((Z(0)^T Z(0))^{-1}) \quad (53)$$

$$= \frac{\|R(t)\|_F^2}{\lambda_r(Z(0)^T Z(0))}$$

Using this, we can write

$$\int_{T_1}^{T_1+T} \mathbb{1}\left[\frac{d\ell(t)}{dt} > -\delta \text{ and } \ell(t) > \epsilon\right] \, dt \leq \int_{0}^{T_1+T} \mathbb{1}\left[\frac{d\ell(t)}{dt} > -\delta \text{ and } \ell(t) > \epsilon\right] \, dt$$

$$\leq \int_{0}^{T_1+T} \mathbb{1}\left[\frac{d\ell(t)}{dt} > -\delta \text{ and } \|R(t)\|_F^2 > \epsilon\lambda_r(Z(0)^T Z(0))\right] \, dt. \quad (54)$$

The expressions in the above integral are Lebesgue integrable as product of bounded integrable functions is integrable (over any finite interval). Here, $\mathbb{1}\left[\frac{d\ell(t)}{dt} > -\delta\right]$, $\mathbb{1}[l(t) > \epsilon]$ and $\mathbb{1}\left[\|R(t)\|_F^2 > \epsilon\lambda_r(Z(0)^T Z(0))\right]$ are integrable by Lemma 15.

We can bound this integral using Lemma 9. This gives us

$$\int_{T_1}^{T_1+T} \mathbb{1}\left[\frac{d\ell(t)}{dt} > -\delta \text{ and } \ell(t) > \epsilon\right] \, dt \leq \int_{0}^{T_1+T} \mathbb{1}\left[\frac{d\ell(t)}{dt} > -\delta \text{ and } \|R(t)\|_F^2 > \epsilon\lambda_r(Z(0)^T Z(0))\right] \, dt \quad (55)$$

$$\leq \frac{r \left(\frac{2\delta}{\epsilon\lambda_r(Z(0)^T Z(0))} - \lambda_r(A(0))\right)^2 \lambda_1 \left(Z(0)^T Z(0)\right)^2}{\delta}. \quad (56)$$

Substituting this in Equation 52, we get

$$\int_{T_1}^{T_1+T} \frac{d\ell(t)}{dt} \, \mathbb{1}\left[\frac{d\ell(t)}{dt} \leq -\delta\right] \, dt \leq -\delta \left(T - \frac{r \left(\frac{2\delta}{\epsilon\lambda_r(Z(0)^T Z(0))} - \lambda_r(A(0))\right)^2 \lambda_1 \left(Z(0)^T Z(0)\right)^2}{\delta}\right). \quad (57)$$

Substituting the bound for T, we get

$$\int_{T_1}^{T_1+T} \frac{d\ell(t)}{dt} \, \mathbb{1}\left[\frac{d\ell(t)}{dt} \leq -\delta\right] \, dt \leq -\delta \left(\frac{\ell(T_1) + r \, \lambda_r(A(0))^2 \, \lambda_1 \left(Z(0)^T Z(0)\right)^2 - \epsilon}{\delta}\right). \quad (58)$$

Combining equations 49, 50 and 58, we get

$$\ell(T_1 + T) \leq \epsilon. \quad (59)$$

$\square$

Lemma 11. *For any $\epsilon > 0$ and*

$$T \geq \frac{24}{\epsilon} \left(\frac{\sigma_1(Y)\sigma_1(C)\sigma_1(Z(0))^4 \sqrt{r \; min(m,n)}}{\sigma_r(Z(0))^5} \right),$$

FA dynamics satisfy

$$\min_{t \leq T} \; \ell(t) \leq \epsilon.$$

Proof. For ease of notation, let us define $\alpha, \beta, \gamma, \phi$ as follows

$$
\begin{aligned}
\alpha &= \frac{2}{\epsilon \lambda_r(Z(0)^T Z(0))}, \\
\beta &= \lambda_r(A(0)), \\
\gamma &= \lambda_1(Z(0)^T Z(0))^2 r, \\
\phi &= \ell(0) + r\lambda_r(A(0))^2 \lambda_1(Z(0)^T Z(0))^2 - \epsilon.
\end{aligned}
$$

Then from Lemma 10 (with T_1 set to 0), we know that for

$$T \geq \frac{(\alpha\delta - \beta)^2 \gamma}{\delta} + \frac{\phi}{\delta},$$

FA dynamics satisfy

$$\min_{t \leq T} \; \ell(t) \leq \epsilon.$$

This holds for all $\delta > 0$. Minimizing the bound on T with respect to δ by setting $\delta = \sqrt{\frac{\beta^2 \gamma + \phi}{\gamma \alpha^2}}$, we get that for

$$T \geq 2\alpha\gamma\sqrt{\beta^2 + \frac{\phi}{\gamma}} - 2\alpha\beta\gamma, \tag{60}$$

FA dynamics satisfy

$$\min_{t \leq T} \; \ell(t) \leq \epsilon.$$

In the rest of the proof, we will bound the RHS of equation 60.

$$
\begin{aligned}
2\alpha\gamma\sqrt{\beta^2 + \frac{\phi}{\gamma}} - 2\alpha\beta\gamma &\leq 2\alpha\gamma\sqrt{\beta^2 + \frac{\phi}{\gamma}} + 2|\alpha\beta\gamma| \\
&= 2\alpha\gamma\sqrt{\beta^2 + \frac{\phi}{\gamma}} + 2\alpha|\beta|\gamma.
\end{aligned}
$$

In the last step, we used the fact that α and γ are non-negative. Substituting for $\alpha, \beta, \gamma, \phi$, we get $2\alpha\gamma\sqrt{\beta^2 + \frac{\phi}{\gamma}} - 2\alpha\beta\gamma$

$$
\begin{aligned}
&\leq \frac{4 \; \lambda_1(Z(0)^T Z(0))^2 \; r}{\epsilon \; \lambda_r(Z(0)^T Z(0))} \left(\sqrt{\lambda_r(A(0))^2 + \frac{\ell(0) + r\lambda_r(A(0))^2 \lambda_1(Z(0)^T Z(0))^2 - \epsilon}{r \; \lambda_1(Z(0)^T Z(0))^2}} + |\lambda_r(A(0))| \right) \\
&\leq \frac{4 \; \lambda_1(Z(0)^T Z(0))^2 \; r}{\epsilon \; \lambda_r(Z(0)^T Z(0))} \left(\sqrt{\lambda_r(A(0))^2 + \frac{\ell(0) + r\lambda_r(A(0))^2 \lambda_1(Z(0)^T Z(0))^2}{r \; \lambda_1(Z(0)^T Z(0))^2}} + |\lambda_r(A(0))| \right) \\
&= \frac{4 \; \sigma_1(Z(0))^4 \; r}{\epsilon \; \sigma_r(Z(0))^2} \left(\sqrt{\lambda_r(A(0))^2 + \frac{\ell(0) + r\lambda_r(A(0))^2 \sigma_1(Z(0))^4}{r \; \sigma_1(Z(0))^4}} + |\lambda_r(A(0))| \right)
\end{aligned}
\tag{61}
$$

Now, we use the following bounds

$$|\lambda_r(A(0))| = \left|\lambda_r\left(\left(Z(0)^T Z(0)\right)^{-1} CW(0)^T + W(0)C^T \left(Z(0)^T Z(0)\right)^{-1}\right)\right|$$
$$\leq \frac{2\sigma_1(C)\sigma_1(W(0))}{\sigma_r(Z(0))^2} \tag{62}$$
$$\leq \frac{2\sigma_1(C)\sigma_1((Y))}{\sigma_r(Z(0))^3}.$$

Here, for the last inequality, we use $W(0) = (Z(0)^T Z(0))^{-1} Z(0)^T Y$ and therefore $\sigma_1(W(0)) \leq \sigma_1((Y))/\sigma_r(Z(0))$.

$$\ell(0) = \left\|\left(Y - \hat{Y}(0)\right) C^T \left(Z(0)^T Z(0)\right)^{-1/2}\right\|_F^2$$
$$\leq \frac{\sigma_1(C)^2}{\sigma_r(Z)^2} \left\|\left(Y - \hat{Y}(0)\right)\right\|_F^2 \tag{63}$$
$$=\leq \frac{\sigma_1(C)^2}{\sigma_r(Z)^2} \left\|\left(Y - Z(0)\left(Z(0)^T Z(0)\right)^{-1} Z(0)^T Y\right)\right\|_F^2$$
$$\leq \frac{\sigma_1(C)^2}{\sigma_r(Z)^2} \|Y\|_F^2$$

Substituting bounds in Equation 62 and 63 to Equation 61, we get $2\alpha\gamma\sqrt{\beta^2 + \frac{\phi}{\gamma}} - 2\alpha\beta\gamma$

$$\leq \frac{8\ r\ \sigma_1(Z(0))^4\ \sigma_1(C)\ \sigma_1(Y)}{\epsilon\ \sigma_r(Z(0))^5}\left(\sqrt{2 + \frac{\|Y\|_F^2\ \sigma_r(Z(0))^4}{4r\sigma_1(Y)^2\ \sigma_1(Z(0))^4}} + 1\right)$$

$$\leq \frac{8\ r\ \sigma_1(Z(0))^4\ \sigma_1(C)\ \sigma_1(Y)}{\epsilon\ \sigma_r(Z(0))^5}\left(\sqrt{2 + \frac{\|Y\|_F^2}{4r\sigma_1(Y)^2}} + 1\right)$$

$$\leq \frac{8\ r\ \sigma_1(Z(0))^4\ \sigma_1(C)\ \sigma_1(Y)}{\epsilon\ \sigma_r(Z(0))^5}\left(\sqrt{2 + \frac{min(m,n)\sigma_1(Y)^2}{4r\sigma_1(Y)^2}} + 1\right)$$

$$= \frac{8\ r\ \sigma_1(Z(0))^4\ \sigma_1(C)\ \sigma_1(Y)}{\epsilon\ \sigma_r(Z(0))^5}\left(\sqrt{2 + \frac{min(m,n)}{4r}} + 1\right)$$

$$\leq \frac{24}{\epsilon}\left(\frac{\sigma_1(Y)\sigma_1(C)\sigma_1(Z(0))^4\sqrt{r\ min(m,n)}}{\sigma_r(Z(0))^5}\right)$$

where we used $min(m,n) \geq r$ for the last inequality. Combining this with Equation 60, we get that for

$$T \geq \frac{24}{\epsilon}\left(\frac{\sigma_1(Y)\sigma_1(C)\sigma_1(Z(0))^4\sqrt{r\ min(m,n)}}{\sigma_r(Z(0))^5}\right),$$

FA dynamics satisfy

$$\min_{t \leq T}\ \ell(t) \leq \epsilon.$$

□

Lemma 12. *For any $\epsilon_1 > 0$ and*

$$T \geq \frac{24}{\epsilon_1}\left(\frac{\sigma_1(Y)\sigma_1(C)\sigma_1(Z(0))^6\sqrt{r\ min(m,n)}}{\sigma_r(Z(0))^5}\right),$$

FA dynamics satisfy

$$\min_{t \leq T}\ \left\|(Y - \hat{Y}(t))C^T\right\|_F^2 \leq \epsilon_1.$$

Proof.

$$\left\| (Y - \hat{Y}(t))C^T \right\|_F^2 = \left\| (Y - \hat{Y}(t))C^T (Z(0)^T Z(0))^{-1/2}(Z(0)^T Z(0))^{1/2} \right\|_F^2$$
$$\leq \ell(t)\, \sigma_1(Z(0))^2$$

Applying Lemma 11 with $\epsilon = \epsilon_1/\sigma_1(Z(0))^2$, we get that for

$$T \geq \frac{24}{\epsilon_1}\left(\frac{\sigma_1(Y)\sigma_1(C)\sigma_1(Z(0))^6 \sqrt{r\, min(m,n)}}{\sigma_r(Z(0))^5} \right),$$

FA dynamics satisfy

$$\min_{t \leq T} \ell(t) \leq \frac{\epsilon_1}{\sigma_1(Z(0))^2},$$

which implies

$$\min_{t \leq T} \left\| (Y - \hat{Y})C^T \right\|_F^2 \leq \epsilon_1.$$

$\square$

This finishes the proof of the first part of Theorem 1 (convergence of the minimum iterate).

Next, we show that $\left\| (Y - \hat{Y}(t))C^T \right\|_F^2$ goes to 0 as t goes to ∞.

Lemma 13. $\left\| (Y - \hat{Y}(t))C^T \right\|_F^2 \to 0$ *as* $t \to \infty$.

Proof. Since $Z(t)^T Z(t) = Z(0)^T Z(0)$ (Fact 3) and since $Z(0)$ is full column rank, it is enough to show that $\ell(t) = \left\| \left(Y - \hat{Y}(t) \right) C^T \left(Z(t)^T Z(t) \right)^{-1/2} \right\|_F^2$ goes to 0 as t goes to ∞.

We will show this by contradiction. Suppose $\ell(t)$ does not converge to 0 with time. Then there must exist some $\epsilon_1 > 0$, such that for all $T \geq 0$, there exists some $t \geq T$ satisfying $\ell(t) > \epsilon_1$.

We also know from Lemma 10, that for all $\epsilon > 0$ and for all $T \geq 0$, there exists some $t \geq T$ satisfying $l(t) \leq \epsilon$.

Using the above two arguments, we can generate an increasing infinite sequence of times $T_1, T_1', T_2, T_2', \cdots$, such that $l(T_i) \leq \epsilon_1/2$ and $l(T_i') > \epsilon_1$, for all $i \in \mathbb{N}$.

By definition of this sequence,

$$\int_{T_i}^{T_i'} \frac{d\ell}{dt}\, dt = \ell(T_i') - \ell(T_i)$$
$$> \frac{\epsilon_1}{2},$$

for all $i \in \mathbb{N}$. Let k be some integer greater than $\frac{2r\, \lambda_r(A(0))^2\, \lambda_1\left(Z(0)^T Z(0)\right)^2}{\epsilon_1}$. Then,

$$\int_0^{T_k'} \frac{d\ell}{dt}\, \mathbb{1}\left[\frac{d\ell}{dt} \geq 0 \right]\, dt \geq \int_0^{T_k'} \frac{d\ell}{dt}\, dt$$
$$\geq \sum_{i=1}^k \int_{T_i}^{T_i'} \frac{d\ell}{dt}\, dt$$
$$> \frac{k\epsilon_1}{2}$$
$$> r\, \lambda_r\left(A(0)\right)^2\, \lambda_1\left(Z(0)^T Z(0)\right)^2.$$

This is a contradiction to Lemmma 8, where we show for all T,

$$\int_0^T \frac{d\ell(t)}{dt}\, \mathbb{1}\left[\frac{d\ell(t)}{dt} \geq 0 \right]\, dt \leq r\, \lambda_r\left(A(0)\right)^2\, \lambda_1\left(Z(0)^T Z(0)\right)^2.$$

Therefore, by contradiction $\ell(t)$ must converge to 0 as t goes to ∞. □

This completes the proof of Theorem 1.

Below, we prove some helper Lemmas that we used to show that the integrals involved in the above proofs were well defined.

Lemma 14. $\ell(t)$, $\frac{d\ell(t)}{dt}$, $\|R(t)\|_F$, $\|A(t)\|_F$ are bounded for all t.

Proof. The proof follows from the fact that $\|Z(t)\|_F$ and $\left\|\left(Z(t)^T Z(t)\right)^{-1}\right\|_F$ are bounded which holds since $Z(t)^T Z(t)$ doesn't change with time (Fact 3). □

Lemma 15. Let $f(t)$ be some continuous function of t over $[0, \infty)$. Then the functions $\mathbb{1}[f(t) > 0]$ and $\mathbb{1}[f(t) \geq 0]$ are Lebesgue integrable over $[T_1, T_2]$ for all $T_1, T_2 \geq 0$.

Proof. Since f is a continuous function, $\{t : f(t) > 0\}$ and $\{t : f(t) \geq 0\}$ are open and closed sets respectively, which imply they are measurable. Therefore, $\mathbb{1}[f(t) > 0]$ and $\mathbb{1}[f(t) \geq 0]$ are bounded measurable functions. The proof follows from the fact that bounded measurable functions over any finite interval are Lebesgue integrable . □

Lemma 16. $f_i(t) = \left(x^T R_{\leq k}^i(t)\right)^2$ and $g_i(t) = \left\|R_{\leq k}^i(t)\right\|^2$ are Lebesgue integrable over $[0, T]$ for all T, k, x, i.

Proof. We can write $f_i(t) = (x^T R^i(t)x)\, \mathbb{1}[R^i(t)^T A(t) R^i(t) \leq k]$ and $g_i(t) = \left\|R^i(t)\right\|^2 \mathbb{1}[R^i(t)^T A(t) R^i(t) \leq k]$. $(x^T R^i(t)x)$, $\left\|R^i(t)\right\|^2$ and $\mathbb{1}[R^i(t)^T A(t) R^i(t) \leq k]$ are bounded (Lemma 14). Also, $(x^T R^i(t)x)$ and $\left\|R^i(t)\right\|^2$ are integrable as they are continuous and $\mathbb{1}[R^i(t)^T A(t) R^i(t) \leq k]$ is integrable by Lemma 16. The Lemma follows since the product of bounded Lebesgue integrable functions is Lebesgue integrable (over any finite interval).

□

B PROOF OF LEMMA 1 AND THEOREM 2

Lemma 1. *Suppose C is chosen randomly with entries drawn i.i.d. from $\mathcal{N}(0,1)$ and the stationary point equations for feedback alignment (7) are satisfied. Let $A_{n \times r} = YC^T$ and $B_{r \times m} = \arg\min_B \|AB - Y\|_F^2$. Then $\hat{Y} = AB$ almost surely.*

Proof. From stationary point equation 7, we know that $(Y - \hat{Y})C^T = 0$, which gives us

$$YC^T = ZWC^T$$

This implies

$$col(A) = col(YC^T) \subseteq col(Z) \tag{64}$$

where $col(\cdot)$ denotes the space given the linear span of columns of the corresponding matrix.

We also know from stationary point equation 7 that $Z^T(Y - ZW) = 0$. This implies that

$$W = \arg\min_W \|ZW - Y\|_F^2 \tag{65}$$

That is, W is chosen optimally once we fix Z.

We consider two cases depending on rank of Y.

Case 1: $rank(Y) \leq r$.

Since $C_{n \times r}$ is a random matrix with i.i.d. Gaussian entries, we get $col(A) = col(YC^T) = col(Y)$ almost surely. Thus B minimizing $\|AB - Y\|_F^2$ will satisfy $AB = Y$ almost surely.

Also, $col(YC^T) = col(Y) \subseteq col(Z)$ almost surely (from equation 64). Thus W minimizing $\|ZW - Y\|_F^2$ will satisfy $ZW = Y$ almost surely. So, we get that $\hat{Y} = AB$ almost surely.

Case 2: $rank(Y) > r$.

Since $C_{n \times r}$ is a random matrix with i.i.d. Gaussian entries and $rank(Y) > r$, $rank(YC^T) = r$ almost surely. We also know that $rank(Z_{n \times r}) \leq r = rank(YC^T)$, and $col(A) = col(YC^T) \subseteq col(Z)$ (from equation 64). This implies $col(A) = col(YC^T) = col(Z)$ and $rank(A) = rank(Z) = r$ almost surely. Therefore, we can write $A = ZR$ for some invertible matrix R, almost surely. Recall that $B = \arg\min_B \|AB - Y\|_F^2$ and $W = \arg\min_B \|AB - Y\|_F^2$. Since A and Z are full column rank, we get $B = (A^T A)^{-1} A^T Y$ and $W = (Z^T Z)^{-1} Z^T Y$, almost surely. Substituting $A = ZR$, we get

$$
\begin{aligned}
AB &= ZR(R^T Z^T Z R)^{-1} R^T Z^T Y \\
&= Z(Z^T Z)^{-1} Z^T Y \\
&= ZW \\
&= \hat{Y}
\end{aligned}
$$

almost surely. This completes the proof. $\qquad\square$

Theorem 2. *Suppose C is chosen randomly with entries drawn i.i.d. from $\mathcal{N}(0, 1)$, the stationary point equations for feedback alignment (7) are satisfied, and $r \geq rank(Y)$. Then $ZW = \hat{Y} = Y$ almost surely, which also minimizes $\|ZW - Y\|_F^2$.*

Proof. In case 1 of the proof of Lemma 1, we show ZW = Y almost surely, when $rank(Y) \leq r$. This proves Theorem 2. $\qquad\square$

C PROOF OF THEOREM 3

For the proofs below, we will use P_X to denote the matrix projecting onto the linear span of columns of matrix X, and $X_{i:j}$ to denote a matrix containing column i to column j from matrix X. We use $M^{(i)}$ to denote the ith column of M.

Before proving Theorem 3, we prove the following helper lemma. Let σ_i, u_i and v_i be the ith singular value, left singular vector and right singular vector of Y respectively, such that $Y = \sum_{i=1}^n \sigma_i u_i v_i^T = U \Sigma V^T$.

Lemma 17. *Let A be some $n \times r$ matrix and and P_A be the projection matrix for the columns space of A. Let $B = \arg\min_B \|Y - AB\|_F^2$. Then $\|Y - AB\|_F^2 = \sum_{i=1}^n \sigma_i^2 (1 - \|P_A u_i\|_2^2)$.*

Proof. Let the singular value decomposition of $A = U_A \Sigma_A V_A^T$ where U_A, Σ_A and V_A are $n \times rank(A)$, $rank(A) \times rank(A)$ and $rank(A) \times r$ matrices respectively. Since $B = \arg\min_B \|Y - AB\|_F^2$, we know

$$
\begin{aligned}
AB &= A A^+ Y \\
&= U_A U_A^T Y
\end{aligned}
$$

. Here A^+ denotes the pseudoinverse of A. This gives us

$$
\begin{aligned}
\|Y - AB\|_F^2 &= \left\|Y - U_A U_A^T Y\right\|_F^2 \\
&= \|Y\|_F^2 + Tr(Y^T U_A U_A^T U_A U_A^T Y) - 2Tr(Y^T U_A U_A^T Y) \\
&= \|Y\|_F^2 - \left\|U_A^T Y\right\|_F^2 \\
&= \sum_{i=1}^n \sigma_i^2 - \sum_{i=1}^n \sigma_i^2 \left\|U_A^T u_i\right\|_2^2 \\
&= \sum_{i=1}^n \sigma_i^2 - \sum_{i=1}^n \sigma_i^2 \left\|U_A U_A^T u_i\right\|_2^2 \\
&= \sum_{i=1}^n \sigma_i^2 (1 - \|P_A u_i\|_2^2).
\end{aligned}
$$

This completes the proof. $\qquad\square$

Theorem 3. *Suppose*
(i) C has entries drawn i.i.d. from $\mathcal{N}(0,1)$,
(ii) the stationary point equations for feedback alignment (7) are satisfied,
(iii) the singular values of Y satisfy

$$
\sigma_i = \begin{cases} \frac{1}{\sqrt{2r}}, & for \ \ i \leq r \\ \frac{1}{\sqrt{2(n-r)}}, & for \ \ r+1 \leq i \leq n \end{cases}
$$

(iv) $c_1 \leq r \leq c_2 n$ for some absolute constants c_1, c_2,

then the error $\|ZW - Y\|_F^2 \geq 0.74$ with probability at least 0.99 over the choice of C.

On the other hand, gradient flow (2) starting from randomly initialized Z and W (with i.i.d. $\mathcal{N}(0,1)$ entries) converges to the optimum solution with $\|ZW - Y\|_F^2 = 0.5$ almost surely.

Proof. The part about gradient flow follows from prior results. From Bah et al. (2019, Theorem 39, part (b)) we know that gradient flow starting from randomly initialized Z and W reaches the global optimum almost surely. The global optimum here corresponds to the best rank r approximation of Y (in Frobenius norm) (Blum et al., 2020) whose error $\|ZW - Y\|_F^2$ is given by $\sum_{i=r+1}^n \sigma_i^2 = 0.5$.

Next, we lower bound the error at Z and W satisfying the stationary point equations for feedback alignment. From Lemma 1, we know that $ZW = AB$ almost surely, where $A = YC^T$ and $B = \arg\min_B \|AB - Y\|_F^2$. Using Lemma 17, this gives us

$$
\begin{aligned}
\|ZW - Y\|_F^2 &= \|AB - Y\|_F^2 \\
&= \sum_{i=1}^n \sigma_i^2 (1 - \|P_A u_i\|_2^2) \\
&= 1 - \sum_{i=1}^n \sigma_i^2 \|P_A u_i\|_2^2,
\end{aligned} \tag{66}
$$

almost surely. We used $\sum_{i=1}^n \sigma_i^2 = 1$ in the last step.

For the rest of the proof, we will assume that $A = YC^T$ is full column rank. Since Y has rank n, and $C_{n \times r}$ is a random matrix with $r \leq n$, this is true almost surely over the choice of C. Since we want to bound the error with high probability (at least 0.9), we can safely ignore the cases where A is not full column rank.

We want to lower bound $\|ZW - Y\|_F^2$, for which we will upper bound $\sum_{i=1}^n \sigma_i^2 \|P_A u_i\|_2^2$. Recall that $A = YC^T = \sum_{i=1}^n \sigma_i u_i v_i^T C$. We can write the jth column of A, $A^{(j)} = \sum_{i=1}^n \sigma_i u_i R_{ij}$, where $R_{ij} = \langle v_i, C^{(j)} \rangle$. Since

the entries of C are drawn i.i.d. from $\mathcal{N}(0,1)$ and v_is are orthonormal, R_{ij}s are $\mathcal{N}(0,1)$ random variables, and are independent for all i,j.

Let $\hat{A}$ be the matrix obtained by applying the Gram–Schmidt orthonormalization process to the columns of A. That is, $\hat{A}^{(1)} = \frac{A^{(1)}}{\|\hat{A}^{(1)}\|}$ and $\hat{A}^{(k)} = \frac{A^{(k)} - P_{A_{1:k-1}}A^{(k)}}{\|A^{(k)} - P_{A_{1:k-1}}A^{(k)}\|}$ for $2 \leq k \leq r$. Here $P_{A_{1:k-1}}$ is the projection matrix for projecting to the space spanned by the first $k-1$ columns of A. We can write $P_A = \hat{A}\hat{A}^T$, which gives

$$
\begin{aligned}
\sum_{i=1}^{n} \sigma_i^2 \|P_A u_i\|_2^2 &= \sum_{i=1}^{n} \sigma_i^2 \left\|\hat{A}^T u_i\right\|_2^2 \\
&= \sum_{i=1}^{r} \sigma_i^2 \left\|\hat{A}^T u_i\right\|_2^2 + \sum_{i=r+1}^{n} \sigma_i^2 \left\|\hat{A}^T u_i\right\|_2^2 \\
&= \frac{1}{2r} \sum_{i=1}^{r} \left\|\hat{A}^T u_i\right\|_2^2 + \frac{1}{2(n-r)} \sum_{i=r+1}^{n} \left\|\hat{A}^T u_i\right\|_2^2 .
\end{aligned}
\tag{67}
$$

Since u_is are orthonormal, this gives us

$$
\sum_{i=r+1}^{n} \left\|\hat{A}^T u_i\right\|_2^2 \leq \left\|\hat{A}\right\|_F^2
$$

$$
\leq r
$$

where we used the fact that $\hat{A}$ is an $n \times r$ matrix with unit length columns. Therefore we get

$$
\sum_{i=1}^{n} \sigma_i^2 \|P_A u_i\|_2^2 \leq \frac{1}{2r} \sum_{i=1}^{r} \left\|\hat{A}^T u_i\right\|_2^2 + \frac{r}{2(n-r)}
\tag{68}
$$

Now we need to bound $\sum_{i=1}^{r} \left\|\hat{A}^T u_i\right\|_2^2$. Let $U_{n \times n}$ be a matrix whose j^{th} column $U^{(j)} = u_j$ and let $U_{i:j}$ be a matrix with columns $u_i, u_{i+1} \cdots, u_j$. With this notation, we need to bound $\left\|\hat{A}^T U_{1:r}\right\|_F^2 = \sum_{j=1}^{r} \left\|U_{1:r}^T \hat{A}^{(j)}\right\|_2^2$. To get a sense of this quantity, we first consider $\left\|U_{1:r}^T A^{(j)}\right\|_2^2$.

$$
\begin{aligned}
\left\|U_{1:r}^T A^{(j)}\right\|_2^2 &= \left\|U_{1:r}^T \left(\sum_{i=1}^{n} \sigma_i u_i R_{ij}\right)\right\|_2^2 \\
&= \sum_{i=1}^{r} \sigma_i^2 \|u_i\|_2^2 R_{ij}^2 \\
&= \frac{1}{2r} \sum_{i=1}^{r} R_{ij}^2
\end{aligned}
$$

which is at most $\frac{1}{2}$ plus some lower order term with high probability. We will show that $\left\|U_{1:r}^T \hat{A}^{(j)}\right\|_2^2$ also can not be much larger than $\frac{1}{2}$ with high probability. To show that, we consider the ratio $\frac{\|U_{1:r}^T \hat{A}^{(j)}\|_2^2}{\|U_{r+1:n}^T \hat{A}^{(j)}\|_2^2}$. Let $P_{U_{1:r}} = U_{1:r}U_{1:r}^T$ and $P_{U_{r+1:n}} = U_{r+1:n}U_{r+1:n}^T$ be the projection matrices for the space spanned by columns of $U_{1:r}$ and $U_{r+1:n}$ respectively. We know

$$
\begin{aligned}
\frac{\left\|U_{1:r}^T \hat{A}^{(j)}\right\|_2^2}{\left\|U_{r+1:n}^T \hat{A}^{(j)}\right\|_2^2} &= \frac{\left\|P_{U_{1:r}} \hat{A}^{(j)}\right\|_2^2}{\left\|P_{U_{r+1:n}} \hat{A}^{(j)}\right\|_2^2} \\
&= \frac{\left\|P_{U_{1:r}}\left(A^{(j)} - P_{A_{1:j-1}}A^{(j)}\right)\right\|_2^2}{\left\|P_{U_{r+1:n}}\left(A^{(j)} - P_{A_{1:j-1}}A^{(j)}\right)\right\|_2^2}
\end{aligned}
$$

for $j \in \{2, 3, \cdots, r\}$ and

$$\frac{\left\|U_{1:r}^T \hat{A}^{(1)}\right\|_2^2}{\left\|U_{r+1:n}^T \hat{A}^{(1)}\right\|_2^2} = \frac{\left\|P_{U_{1:r}} \hat{A}^{(1)}\right\|_2^2}{\left\|P_{U_{r+1:n}} \hat{A}^{(1)}\right\|_2^2}$$

$$= \frac{\left\|P_{U_{1:r}} A^{(1)}\right\|_2^2}{\left\|P_{U_{r+1:n}} A^{(1)}\right\|_2^2}$$

. Using Lemma 18 and 19, and applying a union bound, we get

$$\frac{\left\|U_{1:r}^T \hat{A}^{(j)}\right\|_2^2}{\left\|U_{r+1:n}^T \hat{A}^{(j)}\right\|_2^2} \le \frac{\frac{1}{2} + 4\sqrt{\frac{r}{n-r}}}{\frac{1}{2} - 7\sqrt{\frac{r}{n-r}}}$$

with probability at least 0.99, for all $j \in \{1, 2 \cdots, r\}$ (assuming $r \ge 2000$ and $n \ge 2r$). Since $\hat{A}^{(j)}$ is a unit vector, $\left\|U_{r+1:n}^T \hat{A}^{(j)}\right\|_2^2 = 1 - \left\|U_{1:r}^T \hat{A}^{(j)}\right\|_2^2$. This gives us

$$\frac{\left\|U_{1:r}^T \hat{A}^{(j)}\right\|_2^2}{1 - \left\|U_{1:r}^T \hat{A}^{(j)}\right\|_2^2} \le \frac{\frac{1}{2} + 4\sqrt{\frac{r}{n-r}}}{\frac{1}{2} - 7\sqrt{\frac{r}{n-r}}}$$

$$\le \frac{\frac{1}{2} + 7\sqrt{\frac{r}{n-r}}}{\frac{1}{2} - 7\sqrt{\frac{r}{n-r}}}$$

with probability at least 0.99, for all $j \in \{1, 2 \cdots, r\}$. Rearranging, we get

$$\left\|U_{1:r}^T \hat{A}^{(j)}\right\|_2^2 \le \frac{1}{2} + 7\sqrt{\frac{r}{n-r}}$$

with probability at least 0.99, for all $j \in \{1, 2 \cdots, r\}$. This gives us the desired bound on $\sum_{i=1}^r \left\|\hat{A}^T u_i\right\|_2^2$.

$$\sum_{i=1}^r \left\|\hat{A}^T u_i\right\|_2^2 = \sum_{i=1}^r \left\|U_{1:r}^T \hat{A}^{(j)}\right\|_2^2$$

$$\le r\left(\frac{1}{2} + 7\sqrt{\frac{r}{n-r}}\right)$$

(69)

with probability at least 0.99. Combining this with Equation 68, we get

$$\sum_{i=1}^n \sigma_i^2 \|P_A u_i\|_2^2 \le \frac{1}{2r} \sum_{i=1}^r \left\|\hat{A}^T u_i\right\|_2^2 + \frac{r}{2(n-r)}$$

$$\le \frac{1}{4} + \frac{7}{2}\sqrt{\frac{r}{n-r}} + \frac{r}{2(n-r)}$$

$$\le \frac{1}{4} + 4\sqrt{\frac{r}{n-r}}$$

$$\le \frac{1}{4} + 0.01$$

(70)

with probability at least 0.99 (assume $n \ge 40001\, r$). Combining this with equation 66, we get

$$\|ZW - Y\|_F^2 = 1 - \sum_{i=1}^n \sigma_i^2 \|P_A u_i\|_2^2$$

$$\ge 0.74$$

(71)

with probability at least 0.99, assuming $2000 \le r \le \frac{1}{40001}n$. □

We use Lemma 18, 19 and 20 for proving Theorem 3. These Lemmas are proved assuming the conditions (i) to (iv) in Theorem 3 are satisfied.

Lemma 18. *Let $r \geq 2000$ and $n \geq 2r$. With probability at least 0.997,*

$$\left\| P_{U_{1:r}} \left(A^{(j)} - P_{A_{1:j-1}} A^{(j)} \right) \right\|_2^2 \leq \frac{1}{2} + 4\sqrt{\frac{r}{n-r}},$$

$$\left\| P_{U_{1:r}} \left(A^{(1)} \right) \right\|_2^2 \leq \frac{1}{2} + 4\sqrt{\frac{r}{n-r}},$$

for all $j \in \{2, 3 \cdots, r\}$.

Proof. For $2 \leq j \leq r$)

$$\left\| P_{U_{1:r}} \left(A^{(j)} - P_{A_{1:j-1}} A^{(j)} \right) \right\|_2^2 = \left\| P_{U_{1:r}} A^{(j)} - P_{U_{1:r}} P_{A_{1:j-1}} A^{(j)} \right\|_2^2$$

$$= \left\| P_{U_{1:r}} A^{(j)} - P_{U_{1:r}} P_{A_{1:j-1}} \left(P_{U_{1:r}} + P_{U_{r+1:n}} \right) A^{(j)} \right\|_2^2$$

$$= \left\| P_{U_{1:r}} A^{(j)} - P_{U_{1:r}} P_{A_{1:j-1}} P_{U_{1:r}} A^{(j)} - P_{U_{1:r}} P_{A_{1:j-1}} P_{U_{r+1:n}} A^{(j)} \right\|_2^2 \qquad (72)$$

$$\leq \left\| P_{U_{1:r}} A^{(j)} - P_{U_{1:r}} P_{A_{1:j-1}} P_{U_{1:r}} A^{(j)} \right\|_2^2 + \left\| P_{U_{1:r}} P_{A_{1:j-1}} P_{U_{r+1:n}} A^{(j)} \right\|_2^2$$

$$+ 2 \left\| P_{U_{1:r}} A^{(j)} - P_{U_{1:r}} P_{A_{1:j-1}} P_{U_{1:r}} A^{(j)} \right\|_2 \left\| P_{U_{1:r}} P_{A_{1:j-1}} P_{U_{r+1:n}} A^{(j)} \right\|_2$$

Now we bound the individual terms in the last inequality.

$$\left\| P_{U_{1:r}} A^{(j)} - P_{U_{1:r}} P_{A_{1:j-1}} P_{U_{1:r}} A^{(j)} \right\|_2^2 = \left\| P_{U_{1:r}} A^{(j)} \right\|_2^2 + \left\| P_{U_{1:r}} P_{A_{1:j-1}} P_{U_{1:r}} A^{(j)} \right\|_2^2$$

$$- 2 \langle P_{U_{1:r}} A^{(j)}, P_{U_{1:r}} P_{A_{1:j-1}} P_{U_{1:r}} A^{(j)} \rangle$$

Using $\left\| P_{U_{1:r}} P_{A_{1:j-1}} P_{U_{1:r}} A^{(j)} \right\|_2^2 \leq \left\| P_{A_{1:j-1}} P_{U_{1:r}} A^{(j)} \right\|_2^2$ and
$\langle P_{U_{1:r}} A^{(j)}, P_{U_{1:r}} P_{A_{1:j-1}} P_{U_{1:r}} A^{(j)} \rangle = \left\| P_{A_{1:j-1}} P_{U_{1:r}} A^{(j)} \right\|_2^2$, we get

$$\left\| P_{U_{1:r}} A^{(j)} - P_{U_{1:r}} P_{A_{1:j-1}} P_{U_{1:r}} A^{(j)} \right\|_2^2 \leq \left\| P_{U_{1:r}} A^{(j)} \right\|_2^2$$

Substituting $A^{(j)} = \sum_{i=1}^n \sigma_i u_i R_{ij}$ and $P_{U_{1:r}} = U_{1:r} U_{1:r}^T$, we get

$$\left\| P_{U_{1:r}} A^{(j)} - P_{U_{1:r}} P_{A_{1:j-1}} P_{U_{1:r}} A^{(j)} \right\|_2^2 \leq \sum_{i=1}^r \sigma_i^2 R_{ij}^2$$

$$= \frac{1}{2r} \sum_{i=1}^r R_{ij}^2$$

Using concentration (Lemma 21) and a union bound, we know

$$Pr \left[\sum_{i=1}^r R_{ij}^2 \geq r + 16\sqrt{r \, log(r)} \right] \leq \frac{2}{r} \leq \frac{1}{1000} \qquad (73)$$

for all $j \in \{1, 2, \cdots, r\}$, for $r \geq 2000$. Therefore we get

$$\left\| P_{U_{1:r}} A^{(j)} - P_{U_{1:r}} P_{A_{1:j-1}} P_{U_{1:r}} A^{(j)} \right\|_2^2 \leq \frac{1}{2} + \frac{8\sqrt{log(r)}}{\sqrt{r}} \qquad (74)$$

with probability at least 0.999 for all $j \in \{2, 3 \cdots, r\}$. Next, we bound the $\left\| P_{U_{1:r}} P_{A_{1:j-1}} P_{U_{r+1:n}} A^{(j)} \right\|_2^2$ term in Equation 72.

$$\left\| P_{U_{1:r}} P_{A_{1:j-1}} P_{U_{r+1:n}} A^{(j)} \right\|_2^2 \leq \left\| P_{A_{1:j-1}} P_{U_{r+1:n}} A^{(j)} \right\|_2^2$$

$$\leq \frac{r}{n-r} \qquad (75)$$

with probability at least 0.999, for all $j \in \{2, 3, \cdots, n\}$ (assuming $r \geq 2000$, $n \geq 2r$). We prove the last inequality in Lemma 20.

Combining Equations 72, 74 and 75, and applying a union bound, we get

$$
\begin{aligned}
\left\|P_{U_{1:r}}\left(A^{(j)} - P_{A_{1:j-1}}A^{(j)}\right)\right\|_2^2 &\leq \frac{1}{2} + \frac{8\sqrt{log(r)}}{\sqrt{r}} + \frac{r}{n-r} + 2\sqrt{\frac{r}{n-r}}\sqrt{\frac{1}{2} + \frac{8\sqrt{log(r)}}{\sqrt{r}}} \\
&\leq \frac{1}{2} + \frac{8\sqrt{log(r)}}{\sqrt{r}} + 3\sqrt{\frac{r}{n-r}} \\
&\leq \frac{1}{2} + 4\sqrt{\frac{r}{n-r}}
\end{aligned}
\tag{76}
$$

with probability at least 0.998, for all $j \in \{2, 3, \cdots, r\}$ (assuming $r \geq 2000$ and $n \geq 2r$). For the $j = 1$ case, we know

$$
\begin{aligned}
\left\|P_{U_{1:r}}A^{(1)}\right\|_2^2 &= \sum_{i=1}^{r} \sigma_{i1}^2 R_{i1}^2 \\
&= \frac{1}{2r}\sum_{i=1}^{r} R_{i1}{}^2 \\
&\leq \frac{1}{2} + \frac{8\sqrt{log(r)}}{\sqrt{r}} \\
&\leq \frac{1}{2} + 4\sqrt{\frac{r}{n-r}}
\end{aligned}
\tag{77}
$$

with probability at least 0.999 (for $r \geq 2000$). Here we used the concentration inequality from Lemma 21. Combining Equations 76 and 77, and applying a union bound, we get

$$
\left\|P_{U_{1:r}}\left(A^{(j)} - P_{A_{1:j-1}}A^{(j)}\right)\right\|_2^2 \leq \frac{1}{2} + 4\sqrt{\frac{r}{n-r}}
$$

$$
\left\|P_{U_{1:r}}\left(A^{(1)}\right)\right\|_2^2 \leq \frac{1}{2} + 4\sqrt{\frac{r}{n-r}}
$$

with probability at least 0.997, for all $j \in \{2, 3 \cdots, r\}$ (assuming $r \geq 2000$ and $n \geq 2r$). $\qquad \square$

Lemma 19. *Let $r \geq 2000$ and $n \geq 2r$. With probability at least 0.993,*

$$
\left\|P_{U_{r+1:n}}\left(A^{(j)} - P_{A_{1:j-1}}A^{(j)}\right)\right\|_2^2 \geq \frac{1}{2} - \frac{7\sqrt{r}}{\sqrt{n-r}},
$$

$$
\left\|P_{U_{r+1:n}}\left(A^{(1)}\right)\right\|_2^2 \geq \frac{1}{2} - \frac{7\sqrt{r}}{\sqrt{n-r}}
$$

for all $j \in \{2, 3 \cdots, r\}$.

Proof. For $j \in \{2, 3 \cdots, r\}$,

$$\left\|P_{U_{r+1:n}}\left(A^{(j)} - P_{A_{1:j-1}}A^{(j)}\right)\right\|_2^2 = \left\|P_{U_{r+1:n}}A^{(j)} - P_{U_{r+1:n}}P_{A_{1:j-1}}A^{(j)}\right\|_2^2$$

$$= \left\|P_{U_{r+1:n}}A^{(j)} - P_{U_{r+1:n}}P_{A_{1:j-1}}\left(P_{U_{1:r}} + P_{U_{r+1:n}}\right)A^{(j)}\right\|_2^2$$

$$= \left\|P_{U_{r+1:n}}A^{(j)} - P_{U_{r+1:n}}P_{A_{1:j-1}}P_{U_{1:r}}A^{(j)} - P_{U_{r+1:n}}P_{A_{1:j-1}}P_{U_{r+1:n}}A^{(j)}\right\|_2^2$$

$$\geq \left\|P_{U_{r+1:n}}A^{(j)} - P_{U_{r+1:n}}P_{A_{1:j-1}}P_{U_{1:r}}A^{(j)}\right\|_2^2 + \left\|P_{U_{r+1:n}}P_{A_{1:j-1}}P_{U_{r+1:n}}A^{(j)}\right\|_2^2$$

$$- 2\left\|P_{U_{r+1:n}}A^{(j)} - P_{U_{r+1:n}}P_{A_{1:j-1}}P_{U_{1:r}}A^{(j)}\right\|_2 \left\|P_{U_{r+1:n}}P_{A_{1:j-1}}P_{U_{r+1:n}}A^{(j)}\right\|_2$$

$$\geq \left\|P_{U_{r+1:n}}A^{(j)} - P_{U_{r+1:n}}P_{A_{1:j-1}}P_{U_{1:r}}A^{(j)}\right\|_2^2$$

$$- 2\left\|P_{U_{r+1:n}}A^{(j)} - P_{U_{r+1:n}}P_{A_{1:j-1}}P_{U_{1:r}}A^{(j)}\right\|_2 \left\|P_{U_{r+1:n}}P_{A_{1:j-1}}P_{U_{r+1:n}}A^{(j)}\right\|_2 \tag{78}$$

Now we bound the individual terms in the last inequality. We first lower bound the $\left\|P_{U_{r+1:n}}A^{(j)} - P_{U_{r+1:n}}P_{A_{1:j-1}}P_{U_{1:r}}A^{(j)}\right\|_2^2$ term.

$$\left\|P_{U_{r+1:n}}A^{(j)} - P_{U_{r+1:n}}P_{A_{1:j-1}}P_{U_{1:r}}A^{(j)}\right\|_2^2 \geq \left\|P_{U_{r+1:n}}A^{(j)}\right\|_2^2 - 2\langle P_{U_{r+1:n}}A^{(j)}, P_{U_{r+1:n}}P_{A_{1:j-1}}P_{U_{1:r}}A^{(j)}\rangle \tag{79}$$

We lower bound the $\left\|P_{U_{r+1:n}}A^{(j)}\right\|_2^2$ term:

$$\left\|P_{U_{r+1:n}}A^{(j)}\right\|_2^2 = \sum_{i=r+1}^{n} \sigma_i^2 R_{ij}^2$$

$$= \frac{1}{2(n-r)} \sum_{i=r+1}^{n} R_{ij}^2 \tag{80}$$

$$\geq \frac{1}{2} - \frac{8\sqrt{\log(r)}}{\sqrt{n-r}}$$

with probability at least 0.999, for all $j \in \{1, 2 \cdots, r\}$. Here we used concentration of chi-squared random variables (Lemma 21) and a union bound. We also assumed $r \geq 2000$ and $n \geq 2r$.

Next, we upper bound $\langle P_{U_{r+1:n}}A^{(j)}, P_{U_{r+1:n}}P_{A_{1:j-1}}P_{U_{1:r}}A^{(j)}\rangle$.

$$\langle P_{U_{r+1:n}}A^{(j)}, P_{U_{r+1:n}}P_{A_{1:j-1}}P_{U_{1:r}}A^{(j)}\rangle = \langle P_{U_{r+1:n}}A^{(j)}, P_{A_{1:j-1}}P_{U_{1:r}}A^{(j)}\rangle$$

$$= \left\langle \sum_{i=r+1}^{n} u_i\sigma_i R_{ij}, P_{A_{1:j-1}}\left(\sum_{i=1}^{r} u_i\sigma_i R_{ij}\right)\right\rangle$$

$$= \frac{\left\|P_{A_{1:j-1}}\left(\sum_{i=1}^{r} u_i R_{ij}\right)\right\|_2}{2\sqrt{r(n-r)}} \left\langle \sum_{i=r+1}^{n} u_i R_{ij}, \frac{P_{A_{1:j-1}}\left(\sum_{i=1}^{r} u_i R_{ij}\right)}{\left\|P_{A_{1:j-1}}\left(\sum_{i=1}^{r} u_i R_{ij}\right)\right\|_2}\right\rangle \tag{81}$$

Now, note that $\sum_{i=r+1}^{n} u_i R_{ij}$ and $P_{A_{1:j-1}}\left(\sum_{i=1}^{r} u_i R_{ij}\right)$ are independent since the latter only depends on R_{ij} for $i \leq r$ and R_{ik} for $k \leq j-1$. Also, for any fixed $P_{A_{1:j-1}}\left(\sum_{i=1}^{r} u_i R_{ij}\right)$, there exists a unit norm vector x lying in the linear span of $\{u_{r+1}, u_{r+2}, \cdots, u_n\}$, such that

$$\left|\left\langle \sum_{i=r+1}^{n} u_i R_{ij}, \frac{P_{A_{1:j-1}}\left(\sum_{i=1}^{r} u_i R_{ij}\right)}{\left\|P_{A_{1:j-1}}\left(\sum_{i=1}^{r} u_i R_{ij}\right)\right\|_2}\right\rangle\right| \leq \left|\left\langle \sum_{i=r+1}^{n} u_i R_{ij}, x\right\rangle\right| \tag{82}$$

for all values of R_{ij}s. Therefore, we can write

$$Pr\left[\left|\left\langle \sum_{i=r+1}^{n} u_i R_{ij}, \frac{P_{A_{1:j-1}}\left(\sum_{i=1}^{r} u_i R_{ij}\right)}{\left\|P_{A_{1:j-1}}\left(\sum_{i=1}^{r} u_i R_{ij}\right)\right\|_2}\right\rangle\right| \geq \alpha\right] \leq Pr\left[\left|\left\langle \sum_{i=r+1}^{n} u_i R_{ij}, x\right\rangle\right| \geq \alpha\right] \tag{83}$$

for all $\alpha > 0$, where the probability is over the randomness of $\sum_{i=r+1}^{n} u_i R_{ij}$ and $P_{A_{1:j-1}}(\sum_{i=1}^{r} u_i R_{ij})$ is fixed. Now, since $\sum_{i=r+1}^{n} u_i R_{ij}$ is an isotropic Gaussian random variable in the space spanned by $\{u_{r+1}, u_{r+2}, \cdots, u_n\}$, we know that $\langle \sum_{i=r+1}^{n} u_i R_{ij}, x \rangle$ and $\langle \sum_{i=r+1}^{n} u_i R_{ij}, u_{r+1} \rangle$ are equal in distribution. Here we used the fact that x is a unit vector lying in the span of $\{u_{r+1}, u_{r+2}, \cdots, u_n\}$. Therefore

$$Pr\left[\left|\left\langle \sum_{i=r+1}^{n} u_i R_{ij}, x \right\rangle\right| \geq \alpha\right] = Pr\left[\left|\left\langle \sum_{i=r+1}^{n} u_i R_{ij}, u_{r+1} \right\rangle\right| \geq \alpha\right]$$
$$= Pr\left[|R_{r+1\ j}| \geq \alpha\right].$$

Combining this with Equation 83, and setting $\alpha = 2\sqrt{log(r)}$, we get

$$Pr\left[\left|\left\langle \sum_{i=r+1}^{n} u_i R_{ij}, \frac{P_{A_{1:j-1}}\left(\sum_{i=1}^{r} u_i R_{ij}\right)}{\left\|P_{A_{1:j-1}}\left(\sum_{i=1}^{r} u_i R_{ij}\right)\right\|_2} \right\rangle\right| \geq 2\sqrt{log(r)}\right] \leq Pr\left[|R_{r+1\ j}| \geq 2\sqrt{log(r)}\right]$$

Using a standard tail bound for Normal random variables (Wainwright, 2015), and a union bound, we get

$$\left|\left\langle \sum_{i=r+1}^{n} u_i R_{ij}, \frac{P_{A_{1:j-1}}\left(\sum_{i=1}^{r} u_i R_{ij}\right)}{\left\|P_{A_{1:j-1}}\left(\sum_{i=1}^{r} u_i R_{ij}\right)\right\|_2} \right\rangle\right| \leq 2\sqrt{log(r)} \qquad (84)$$

with probability at least 0.999 for all $j \in \{2, 3, \cdots, r\}$ where the probability is over $\sum_{i=r+1}^{n} u_i R_{ij}$. But since this holds for any fixed $P_{A_{1:j-1}}(\sum_{i=1}^{r} u_i R_{ij})$, this holds even when we take probability over both $\sum_{i=r+1}^{n} u_i R_{ij}$ and $P_{A_{1:j-1}}(\sum_{i=1}^{r} u_i R_{ij})$. Also,

$$\left\|P_{A_{1:j-1}}\left(\sum_{i=1}^{r} u_i R_{ij}\right)\right\|_2^2 \leq \left\|\sum_{i=1}^{r} u_i R_{ij}\right\|_2^2$$
$$= \sum_{i=1}^{r} R_{ij}^2 \qquad (85)$$
$$\leq r + 16\sqrt{r\ log(r)}$$

with probability at least 0.999, for all $j \in \{2, 3 \cdots, r\}$ (assuming $r \geq 2000$). Here we used concentration (Lemma 21) and a union bound. Combining Equations 81, 84 and 85, and using a union bound, we get

$$\langle P_{U_{r+1:n}} A^{(j)}, P_{U_{r+1:n}} P_{A_{1:j-1}} P_{U_{1:r}} A^{(j)} \rangle \leq \sqrt{\frac{\left(r + 16\sqrt{r\ log(r)}\right)\ 4\ log(r)}{2r(n-r)}}$$
$$\leq 2\sqrt{\frac{log(r)}{n-r}} \qquad (86)$$

with probability at least 0.998, for all $j \in \{2, 3, \cdots, r\}$, assuming $r \geq 2000$. Combining equations 79, 80 and 86, and applying a union bound, we get

$$\left\|P_{U_{r+1:n}} A^{(j)} - P_{U_{r+1:n}} P_{A_{1:j-1}} P_{U_{1:r}} A^{(j)}\right\|_2^2 \geq \frac{1}{2} - \frac{8\sqrt{log(r)}}{\sqrt{n-r}} - \frac{4\sqrt{log(r)}}{\sqrt{n-r}}$$
$$= \frac{1}{2} - \frac{12\sqrt{log(r)}}{\sqrt{n-r}} \qquad (87)$$

with probability at least 0.997, for all $j \in \{2, 3 \cdots, r\}$. We also upper bound

$$\left\| P_{U_{r+1:n}} A^{(j)} - P_{U_{r+1:n}} P_{A_{1:j-1}} P_{U_{1:r}} A^{(j)} \right\|_2^2.$$

$$
\begin{aligned}
\left\| P_{U_{r+1:n}} A^{(j)} - P_{U_{r+1:n}} P_{A_{1:j-1}} P_{U_{1:r}} A^{(j)} \right\|_2^2 &\leq \left\| P_{U_{r+1:n}} A^{(j)} \right\|^2 + \left\| P_{U_{r+1:n}} P_{A_{1:j-1}} P_{U_{1:r}} A^{(j)} \right\|^2 \\
&\quad + 2 \left\| P_{U_{r+1:n}} A^{(j)} \right\| \left\| P_{U_{r+1:n}} P_{A_{1:j-1}} P_{U_{1:r}} A^{(j)} \right\| \\
&\leq 4 \left\| A^{(j)} \right\|^2 \\
&= 4 \sum_{i=1}^n \sigma_i^2 R_{ij}^2 \\
&= \frac{2}{r} \sum_{i=1}^r R_{ij}^2 + \frac{2}{n-r} \sum_{i=r+1}^n R_{ij}^2 \\
&\leq 4 + \frac{32 \sqrt{log(r)}}{\sqrt{r}} + \frac{32 \sqrt{log(r)}}{\sqrt{n-r}} \\
&\leq 8
\end{aligned}
\tag{88}
$$

with probability at least 0.998 for all $j \in 2, 3, \cdots, r$ (assuming $r \geq 2000$ and $n \geq 2r$). Here we applied concentration (Lemma 21) and a union bound.

We also need to upper bound $\left\| P_{U_{r+1:n}} P_{A_{1:j-1}} P_{U_{r+1:n}} A^{(j)} \right\|_2$.

$$
\begin{aligned}
\left\| P_{U_{r+1:n}} P_{A_{1:j-1}} P_{U_{r+1:n}} A^{(j)} \right\|_2^2 &\leq \left\| P_{A_{1:j-1}} P_{U_{r+1:n}} A^{(j)} \right\|_2^2 \\
&\leq \frac{r}{n-r}
\end{aligned}
\tag{89}
$$

with probability at least 0.999, for all $j \in \{2, 3, \cdots, n\}$ (assuming $r \geq 2000$, $n \geq 2r$). We prove the last inequality in Lemma 20.

Combining Equations 78, 87, 88 and 89, and applying a union bound, we get

$$
\begin{aligned}
\left\| P_{U_{r+1:n}} \left(A^{(j)} - P_{A_{1:j-1}} A^{(j)} \right) \right\|_2^2 &\geq \frac{1}{2} - \frac{12 \sqrt{log(r)}}{\sqrt{n-r}} - 2\sqrt{8} \sqrt{\frac{r}{n-r}} \\
&\geq \frac{1}{2} - \frac{7\sqrt{r}}{\sqrt{n-r}}
\end{aligned}
\tag{90}
$$

with probability at least 0.994 for all $j \in \{2, 3, \cdots, r\}$ (assuming $r \geq 2000$, $n \geq 2r$).

To prove the lemma, we also need to bound $\left\| P_{U_{r+1:n}} A^{(1)} \right\|_2^2$.

$$
\begin{aligned}
\left\| P_{U_{r+1:n}} A^{(1)} \right\|_2^2 &= \sum_{i=r+1}^n \sigma_i^2 R_{i1}^2 \\
&= \frac{1}{2(n-r)} \sum_{i=r+1}^n R_{i1}^2 \\
&\geq \frac{1}{2} - \frac{16\sqrt{log(r)}}{\sqrt{n-r}} \\
&\geq \frac{1}{2} - \frac{7\sqrt{r}}{\sqrt{n-r}}
\end{aligned}
\tag{91}
$$

with probability at least 0.999 (assuming $r \geq 2000$, $n \geq 2r$). Here we used the concentration inequality from Lemma 21.

Applying a union bound, we get that both Equations 90 and 91 hold with probability at least 0.993 for all $j \in \{2, 3, \cdots, r\}$. This completes the proof. $\qquad\square$

Lemma 20. *Let $r \geq 2000$ and $n \geq 2r$. Then*

$$\left\| P_{A_{1:j-1}} P_{U_{r+1:n}} A^{(j)} \right\|_2^2 \leq \frac{r}{n-r}.$$

with probability at least 0.999 for all $j \in \{2, 3, \cdots, r\}$.

Proof.

$$\left\| P_{A_{1:j-1}} P_{U_{r+1:n}} A^{(j)} \right\|_2^2 \leq \left\| P_{A_{1:j-1}} \left(\sum_{i=r+1}^{n} \sigma_i u_i R_{ij} \right) \right\|_2^2$$

$$= \frac{1}{2(n-r)} \left\| P_{A_{1:j-1}} \left(\sum_{i=r+1}^{n} u_i R_{ij} \right) \right\|_2^2 \tag{92}$$

Note that $P_{A_{1:j-1}}$ and $\left(\sum_{i=r+1}^{n} u_i R_{ij} \right)$ are independent since $P_{A_{1:j-1}}$ only depends on random variables R_{ik} for $k \leq j-1$. Also, for any fixed $P_{A_{1:j-1}}$, there exists a projection matrix P_{j-1} that project onto some $j-1$ dimensional subspace of the linear span of $\{u_{r+1}, u_{r+2}, \cdots, u_n\}$ (assume $n \geq 2r$), such that

$$\left\| P_{A_{1:j-1}} \left(\sum_{i=r+1}^{n} u_i R_{ij} \right) \right\|_2^2 \leq \left\| P_{j-1} \left(\sum_{i=r+1}^{n} u_i R_{ij} \right) \right\|_2^2$$

for all values of R_{ij}s. Therefore, we can write

$$Pr \left[\left\| P_{A_{1:j-1}} \left(\sum_{i=r+1}^{n} u_i R_{ij} \right) \right\|_2^2 \geq \alpha \right] \leq Pr \left[\left\| P_{j-1} \left(\sum_{i=r+1}^{n} u_i R_{ij} \right) \right\|_2^2 \geq \alpha \right] \tag{93}$$

for all $\alpha > 0$, where the probability is over the randomness of R_{ij}s and $P_{A_{1:j-1}}$ is fixed. Now, since $\sum_{i=r+1}^{n} u_i R_{ij}$ is an isotropic Gaussian random variable in the space spanned by $\{u_{r+1}, u_{r+2}, \cdots, u_n\}$, we know that $\left\| P_{j-1} \left(\sum_{i=r+1}^{n} u_i R_{ij} \right) \right\|_2^2$ and $\left\| P_{U_{r+1:r+j-1}} \left(\sum_{i=r+1}^{n} u_i R_{ij} \right) \right\|_2^2$ are equal in distribution. Here we used the fact that P_{j-1} projects onto some $j-1$ dimensional subspace of the linear span of $\{u_{r+1}, u_{r+2}, \cdots, u_n\}$. Therefore

$$Pr \left[\left\| P_{j-1} \left(\sum_{i=r+1}^{n} u_i R_{ij} \right) \right\|_2^2 \geq \alpha \right] = Pr \left[\left\| P_{U_{r+1:r+j-1}} \left(\sum_{i=r+1}^{n} u_i R_{ij} \right) \right\|_2^2 \geq \alpha \right]$$

$$= Pr \left[\sum_{i=r+1}^{r+j-1} R_{ij}^2 \geq \alpha \right]$$

$$\leq Pr \left[\sum_{i=r+1}^{2r} R_{ij}^2 \geq \alpha \right]$$

for $j \leq r$. Combining this with Equation 93, and setting $\alpha = r + 16\sqrt{r \ log(r)}$, we get

$$Pr \left[\left\| P_{A_{1:j-1}} \left(\sum_{i=r+1}^{n} u_i R_{ij} \right) \right\|_2^2 \geq r + 16\sqrt{r \ log(r)} \right] \leq Pr \left[\sum_{i=r+1}^{2r} R_{ij}^2 \geq r + 16\sqrt{r \ log(r)} \right]. \tag{94}$$

Using concentration (Lemma 21) and a union bound, we know

$$Pr \left[\sum_{i=r+1}^{2r} R_{ij}^2 \geq r + 16\sqrt{r \ log(r)} \right] \leq \frac{2}{r} \leq \frac{1}{1000}$$

for all $j \in \{1, 2, \cdots, r\}$. Therefore, we get

$$\left\| P_{A_{1:j-1}} \left(\sum_{i=r+1}^{n} u_i R_{ij} \right) \right\|_2^2 \leq r + 16\sqrt{r \ log(r)} \tag{95}$$

with probability at least 0.999 for all $j \in \{2, 3, \cdots, r\}$ where the probability is over R_{ij}s. But since this holds for any fixed $P_{A_{1:j-1}}$, this holds even when we take probability over both R_{ij}s and $P_{A_{1:j-1}}$. Substituting this in Equation 92, we get that for $r \geq 2000$,

$$
\begin{aligned}
\left\| P_{A_{1:j-1}} P_{U_{r+1:n}} A^{(j)} \right\|_2^2 &\leq \frac{r + 16\sqrt{r \, log(r)}}{2(n-r)} \\
&\leq \frac{r}{n-r}.
\end{aligned}
\tag{96}
$$

with probability at least 0.999 for all $j \in \{2, 3, \cdots, r\}$. □

Lemma 21. *(Wainwright, 2015) Let X be a χ^2 random variable with d degrees of freedom, then*

$$
Pr[|X - d| \geq dt] \leq 2e^{-dt^2/8}
$$

for all $t \in (0, 1)$.

D PROOF OF THEOREM 4

Theorem 4. *Suppose*
(i) C has entries drawn i.i.d. from $\mathcal{N}(0, 1)$,
(ii) the stationary point equations for feedback alignment (7) are satisfied,
(iii) the singular values of Y satisfy $\sigma_1 = 1$ and $\sigma_i = \epsilon$ for $i > 1$, where $0 < \epsilon < 1$,
(iv) $r = 1$ (rank 1 approximation) and $n \geq c$ for some absolute constant c.

Let Z_{FA} and W_{FA} denote the Z and W satisfying the above conditions respectively, and Z_{GD} and W_{GD} represent the factors found by gradient flow (2) starting from randomly initialized Z and W (with i.i.d. $\mathcal{N}(0, 1)$ entries). Then

$$
\|Z_{FA}W_{FA} - Y\|_F^2 \leq \|Z_{GD}W_{GD} - Y\|_F^2 \left(1 + \frac{2}{\epsilon^2 n} \right)
$$

and

$$
\left| \left\langle \frac{Z_{FA}}{\|Z_{FA}\|_2}, \frac{Z_{GD}}{\|Z_{GD}\|_2} \right\rangle \right| \leq \frac{4}{\epsilon\sqrt{n}}
$$

with probability at least 0.99 over the choice of C and random initialization of gradient flow.

Proof. Let σ_i, u_i and v_i be the ith singular value, left singular vector and right singular vector of Y respectively, such that $Y = \sum_{i=1}^n \sigma_i u_i v_i^T = U\Sigma V^T$.

From Bah et al. (2019, Theorem 39, part (b)) we know that gradient flow starting from randomly initialized Z and W reaches the global optimum almost surely. The global optimum here corresponds to the best rank 1 approximation of Y (in frobenious norm) (Blum et al., 2020) whose error $\|ZW - Y\|_F^2$ is given by $\sum_{i=2}^n \sigma_i^2 = \epsilon^2(n-1)$. So, from prior work, we know that

$$
\frac{Z_{GD}}{\|Z_{GD}\|_2} = u_1
\tag{97}
$$

and

$$
\|Z_{GD}W_{GD} - Y\|_F^2 = \epsilon^2(n-1)
\tag{98}
$$

almost surely.

Next, we consider Z_{FA} and W_{FA} satisfying the stationary point equations for feedback alignment. From Lemma 1, we know that $Z_{FA}W_{FA} = AB$ almost surely, where $A = YC^T$ and $B = \arg\min_B \|AB - Y\|_F^2$.

Using Lemma 17, this gives us

$$\begin{aligned}
\|Z_{FA}W_{FA} - Y\|_F^2 &= \|AB - Y\|_F^2 \\
&= \sum_{i=1}^{n} \sigma_i^2 (1 - \|P_A u_i\|_2^2) \\
&= \epsilon^2(n-1) + 1 - \sum_{i=1}^{n} \sigma_i^2 \|P_A u_i\|_2^2 \\
&\leq \epsilon^2(n-1) + 1 \\
&\leq \epsilon^2(n-1)\left(1 + \frac{2}{\epsilon^2 n}\right)
\end{aligned} \tag{99}$$

almost surely. Here we assume $n \geq 2$ for the last inequality.

Note that $A = YC^T = \sum_{i=1}^{n} \sigma_i u_i R_i$ where $R_i = \langle v_i, C^T \rangle$ (recall $r = 1$ and therefore C^T is $m \times 1$). Since v_is are othonormal and entries of C are i.i.d. $\mathcal{N}(0,1)$, R_is are $\mathcal{N}(0,1)$ random variables and are independent for all i. Since A and Z_{FA} are $n \times 1$ matrices, B and W_{FA} are $1 \times m$ matrices, and $Z_{FA}W_{FA} = AB$ almost surely, we know that $Z_{FA} = cA$ and $W_{FA} = \frac{1}{c}B$ almost surely, for some non-zero constant c. Therefore $\frac{Z_{FA}}{\|Z_{FA}\|_2} = \frac{A}{\|A\|_2}$ almost surely. Therefore, we get

$$\begin{aligned}
\left\langle \frac{Z_{FA}}{\|Z_{FA}\|_2}, \frac{Z_{GD}}{\|Z_{GD}\|_2} \right\rangle^2 &= \left\langle \frac{A}{\|A\|_2}, u_1 \right\rangle^2 \\
&= \frac{\left(\sum_{i=1}^{n} \sigma_i u_i R_i, u_1\right)^2}{\sum_{i=1^n} \sigma_i^2 R_i^2} \\
&= \frac{R_1^2}{R_1^2 + \epsilon^2 \sum_{i=2}^{n} R_i^2} \\
&\leq \frac{R_1^2}{\epsilon^2 \sum_{i=2}^{n} R_i^2} \\
&\leq \frac{10}{\epsilon^2 n}
\end{aligned} \tag{100}$$

with probability at least 0.99. For the last inequality, we used tail bounds for normal random variable and chi-squared random variable (Lemma 21), and a union bound, which give $R_1^2 \leq 9$ and $\sum_{i=2}^{n} R_i^2 \geq n - 1 - 8\sqrt{n} \geq 0.9n$ (assume $n \geq 10000$) with probability at least 0.99. From Equations 98, 99 and 100, we get

$$\|Z_{FA}W_{FA} - Y\|_F^2 \leq \|Z_{GD}W_{GD} - Y\|_F^2 \left(1 + \frac{2}{\epsilon^2 n}\right)$$

and

$$\left| \left\langle \frac{Z_{FA}}{\|Z_{FA}\|_2}, \frac{Z_{GD}}{\|Z_{GD}\|_2} \right\rangle \right| \leq \frac{4}{\epsilon\sqrt{n}}$$

with probability at least 0.99 (assuming $n \geq 10000$). □

E PROOF IDEA OF Baldi et al. (2018)

Baldi et al. (2018, Theorem 8) show convergence of feedback alignment (FA) for training two layer linear neural networks. Here, we discuss the main idea behind their proof. We describe the idea for the problem of matrix factorization. Recall that the FA update is given by

$$\begin{aligned}
\frac{dZ}{dt} &= (Y - \hat{Y})C^T \\
\frac{dW}{dt} &= Z^T(Y - \hat{Y})
\end{aligned} \tag{101}$$

We want to show that $\left\|(Y - \hat{Y})C^T\right\|_F^2$ converges to 0 with time.

Let

$$V = \frac{1}{2}\left(CW^TWC^T - CY^TZ - Z^TYC^T\right)$$

Observe that

$$\frac{dV}{dt} = -C(Y - \hat{Y})^T(Y - \hat{Y})C^T$$

which implies

$$\frac{d\,Tr(V)}{dt} = -\left\|(Y - \hat{Y})C^T\right\|_F^2 \leq 0.$$

Therefore, $Tr(V)$ is monotonically non-increasing with its rate of decrease given by $\left\|(Y - \hat{Y})C^T\right\|_F^2$.

Now, since CW^TWC^T is PSD, $Tr(CW^TWC^T) \geq 0$. Also, $Tr(CY^TZ) = Tr(Z^TYC^T)$. This gives

$$Tr(V) \geq -Tr\left(CY^TZ\right)$$

Baldi et al. (2018) show that Z remains bounded throughout the dynamics. Therefore, $Tr(V)$ is bounded from below.

Since $Tr(V)$ is monotonically non-increasing with rate of decrease given by $\left\|(Y - \hat{Y})C^T\right\|_F^2$, and it is also bounded from below, $\left\|(Y - \hat{Y})C^T\right\|_F^2$ can not be too large for too long. Baldi et al. (2018) use this observation to show that $\left\|(Y - \hat{Y})C^T\right\|_F^2$ converges to 0.

Note that this proof of convergence for FA is very different from our proof of convergence for FA*. This proof doesn't say much about the dynamics of alignment. In contrast, our proof crucially relies on the phenomenon of alignment and sheds light on how it facilitates convergence.

F SIMULATION DETAILS AND ADDITIONAL PLOTS

Figure 2. We generate Y as $U\Sigma V^T$ where U and V are an $n \times k$ and $m \times k$ independent random matrices respectively with orthonormal columns, and Σ is a $k \times k$ diagonal matrix. For GD and FA, Z and W are initialized as $n \times r$ and $r \times m$ random matrices respectively where each entry is drawn i.i.d. from a normal distribution with mean 0 and standard deviation 0.001. We use the same initial Z and W for FA and GD. For FA*, we use the same initial Z as FA and GD but W is initialized to $(Z^TZ)^{-1}Z^TY$. C is a random $r \times m$ matrix with i.i.d. entries drawn from a normal distribution with mean 0 and standard deviation 1. We use the same C for FA, GD and $FA*$. We use a learning rate of 1 for GD and 0.1 for FA and FA*.

For Figure 2a, $n = m = 500$ and $r = k = 50$. The diagonal entries of Σ are set to $1/\sqrt{50}$. For Figure 2b, $n = m = k = 500$ and $r = 50$. The first 50 diagonal entries of Σ are set to $1/\sqrt{2 * 50}$ and the next 450 diagonal entries are set to $1/\sqrt{2 * 450}$. The diagonal entries for 2b are set in accordance with Theorem 3.

Figure 3a. y is a random 100 dimensional unit vector. Z is initialized as a 100×50 random matrix with entries drawn i.i.d. from a normal distribution with mean 0 and standard deviation 0.001. w is initialized as $(Z^TZ)^{-1}Z^Ty$. c is set to $-w(0)$. We use a learning rate of 0.1.

Figure 3b. We generate Y as $AB^T/\left\|AB^T\right\|_F$ where A and B are 100×99 matrices with with entries drawn i.i.d. from a normal distribution with mean 0 and unit standard deviation. Z is initialized as a 100×99 random matrix with orthonormal columns (such that $Z^TZ = I$), and W is initialized as $(Z^TZ)^{-1}Z^TY$. C is a random 99×100 matrix with i.i.d. entries drawn from a normal distribution with mean 0 and standard deviation 1. We use a learning rate of 0.1. We note that the non-monotonic loss progression is not due to any learning rate issue. $\frac{d\left\|(Y - \hat{Y})C^T\right\|_F^2}{dt}$ does switch from being negative to positive and back many times in the dynamics.

Figure 4 This corresponds to the same setting as Figure 3b. We draw x from uniform distribution over the unit sphere.

Figure 5 This corresponds to the same setting as Figure 3b. We use the same $Z(0), W(0), Y$ and C for FA and FA*. For FA and FA*, we use a learnig rate of 0.1. For FA with larger learning rate for W, we use a learning rate of 0.5 for W update and 0.1 for Z update.
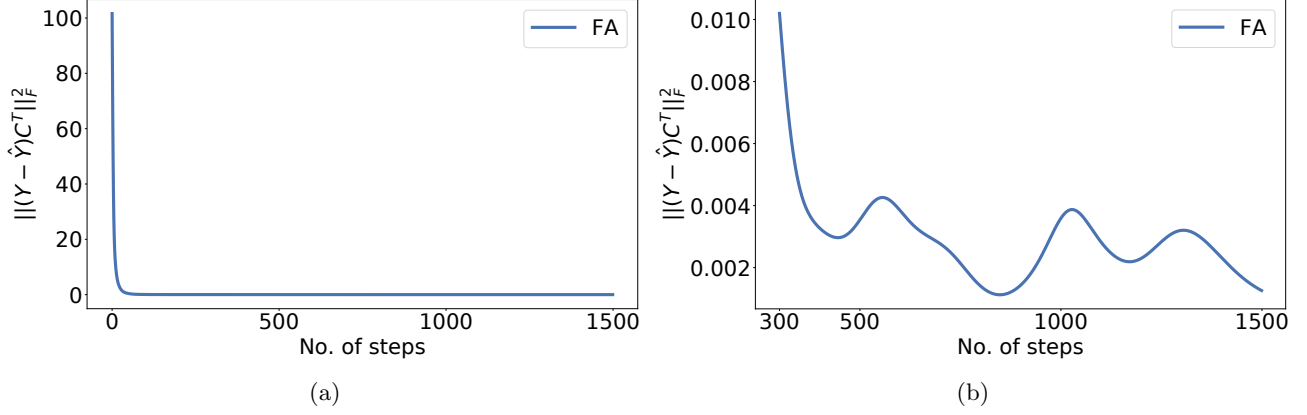

Figure 6: (a) FA loss dynamics with the same Y, C and $Z(0)$ as in Figure 5 but with a randomly initialized W. As expected, FA loss progression in this case looks very different from the loss progression for FA* and for FA with optimally initialized W, as shown in Figure 5. (b) Zoomed in version of Figure 6a starting from step 300 showing non-monotonic loss progression.

Figure 6 Similar to Figure 5, this corresponds to the same setting as Figure 3b, except that W is initialized to a random 99×100 matrix with i.i.d. entries drawn from a normal distribution with mean 0 and standard deviation 0.001.

As expected, the loss progression for FA with randomly initialized W is very different from FA* and from FA with optimally initialized W (shown in Figure 5). The initial loss is much higher in this case compared to Figure 5 as we do not initialize W optimally here.