Convergence and Alignment of Gradient Descent with Random Backpropagation Weights

Ganlin Song* **Ruitu Xu*** **John Lafferty*†**

*Department of Statistics and Data Science
†Wu Tsai Institute
Yale University
{ganlin.song, ruitu.xu, john.lafferty}@yale.edu

Abstract

Stochastic gradient descent with backpropagation is the workhorse of artificial neural networks. It has long been recognized that backpropagation fails to be a biologically plausible algorithm. Fundamentally, it is a non-local procedure—updating one neuron's synaptic weights requires knowledge of synaptic weights or receptive fields of downstream neurons. This limits the use of artificial neural networks as a tool for understanding the biological principles of information processing in the brain. Lillicrap et al. (2016) propose a more biologically plausible "feedback alignment" algorithm that uses random and fixed backpropagation weights, and show promising simulations. In this paper we study the mathematical properties of the feedback alignment procedure by analyzing convergence and alignment for two-layer networks under squared error loss. In the overparameterized setting, we prove that the error converges to zero exponentially fast, and also that regularization is necessary in order for the parameters to become aligned with the random backpropagation weights. Simulations are given that are consistent with this analysis and suggest further generalizations. These results contribute to our understanding of how biologically plausible algorithms might carry out weight learning in a manner different from Hebbian learning, with performance that is comparable with the full non-local backpropagation algorithm.

1 Introduction

The roots of artificial neural networks draw inspiration from networks of biological neurons (Rumelhart et al., 1986a; Elman et al., 1996; Medler, 1998). Grounded in simple abstractions of membrane potentials and firing, neural networks are increasingly being employed as a computational tool for better understanding the biological principles of information processing in the brain; examples include Yildirim et al. (2019) and Yamins & DiCarlo (2016). Even when full biological fidelity is not required, it can be useful to better align the computational abstraction with neuroscience principles.

Stochastic gradient descent has been a workhorse of artificial neural networks. Conveniently, calculation of gradients can be carried out using the backpropagation algorithm, where reverse mode automatic differentiation provides a powerful way of computing the derivatives for general architectures (Rumelhart et al., 1986b). Yet it has long been recognized that backpropagation fails to be a biologically plausible algorithm. Fundamentally, it is a non-local procedure—updating the weight between a presynaptic and postsynaptic neuron requires knowledge of the weights between the postsynaptic neuron and other neurons. No known biological mechanism exists for propagating

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

information in this manner. This limits the use of artificial neural networks as a tool for understanding learning in the brain.

A wide range of approaches have been explored as a potential basis for learning and synaptic plasticity. Hebbian learning is the most fundamental procedure for adjusting weights, where repeated stimulation by a presynaptic neuron that results in the subsequent firing of the postsynapic neuron will result in an increased strength in the connection between the two cells (Hebb, 1961; Paulsen & Sejnowski, 2000). Several variants of Hebbian learning, some making connections to principal components analysis, have been proposed (Oja, 1982; Sejnowski & Tesauro, 1989; Sejnowski, 1999). In this paper, our focus is on a formulation of Lillicrap et al. (2016) based on random backpropagation weights that are fixed during the learning process, called the "feedback alignment" (FA) algorithm. Lillicrap et al. (2016) show that the model can still learn from data, and observe the interesting phenomenon that the error signals propagated with the forward weights align with those propagated with fixed random backward weights during training. Direct feedback alignment (DFA) (Nøkland, 2016) extends FA by adding skip connections to send the error signals directly to each hidden layer, allowing parallelization of weight updates. Empirical studies given by Launay et al. (2020) show that DFA can be successfully applied to train a number of modern deep learning models, including transformers. Based on DFA, Frenkel et al. (2021) proposes direct the random target projection (DRTP) algorithm that trains the network weights with a random projection of the target vector instead of the error, and shows alignment for linear networks. Related proposals, including methods based on the use of differences of neuron activities, have been made in a series of recent papers (Akrout et al., 2019; Bellec et al., 2019; Lillicrap et al., 2020). A comparison of some of these methods is made by Bartunov et al. (2018).

The use of random feedback weights, which are not directly tied to the forward weights, removes issues of non-locality. However, it is not clear under what conditions optimization of error and learning can be successful. While Lillicrap et al. (2016) give suggestive simulations and some analysis for the linear case, it has been an open problem to explain the behavior of this algorithm for training the weights of a neural network. In this paper, we study the mathematical properties of the feedback alignment procedure by analyzing convergence and alignment for two-layer networks under squared error loss. In the overparameterized setting, we prove that the error converges to zero exponentially fast. We also show, unexpectedly, that the parameters become aligned with the random backpropagation weights only when regularization is used. Simulations are given that are consistent with this analysis and suggest further generalizations. The following section gives further background and an overview of our results.

2 Problem Statement and Overview of Results

In this section we provide a formulation of the backpropagation algorithm to establish notation and the context for our analysis. We then formulate the feedback aligment algorithm that uses random backpropation weights. A high-level overview of our results is then presented, together with some of the intuition and proof techniques behind these results; we also contrast with what was known previously.

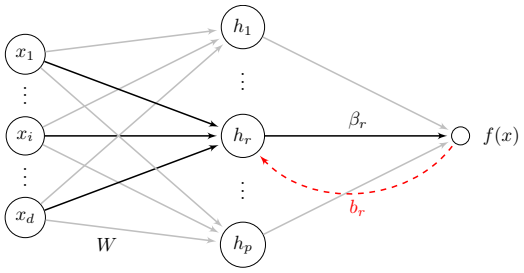
We mainly consider two-layer neural networks in the regression setting, specified by a family of functions $f : \mathbb{R}^d \to \mathbb{R}$ with input dimension d, sample size n, and p neurons in the hidden layer. For an input $x \in \mathbb{R}^d$, the network outputs

$$f(x) = \frac{1}{\sqrt{p}} \sum_{r=1}^{p} \beta_r \psi(w_r^\intercal x) = \frac{1}{\sqrt{p}} \beta^\intercal \psi(Wx), \tag{2.1}$$

where $W = (w_1, ..., w_p)^\intercal \in \mathbb{R}^{p \times d}$ and $\beta = (\beta_1, ..., \beta_p)^\intercal \in \mathbb{R}^p$ represent the feed-forward weights in the first and second layers, and ψ denotes an element-wise activation function. The scaling by $\sqrt{p}$ is simply for convenience in the analysis.

Given n input-response pairs $\{(x_i, y_i)\}_{i=1}^n$, the training objective is to minimize the squared error

$$\mathcal{L}(W, \beta) = \frac{1}{2} \sum_{i=1}^{n} \big(y_i - f(x_i)\big)^2. \tag{2.2}$$



Algorithm 1 Feedback Alignment

Input: Dataset $\{(x_i, y_i)\}_{i=1}^n$, step size η

initialize W, β and b as Gaussian

while not converged **do**

  $\beta_r \leftarrow \beta_r - \frac{\eta}{\sqrt{p}} \sum_{i=1}^n e_i \psi(w_r^\intercal x_i)$

  $w_r \leftarrow w_r - \frac{\eta}{\sqrt{p}} \sum_{i=1}^n e_i b_r \psi'(w_r^\intercal x_i) x_i$

  for $r \in [p]$

end while

Figure 1: Standard backpropagation updates the first layer weights for a hidden node r with the second layer feedforward weight β_r. We study the procedure where the error is backpropagated instead using a fixed, random weight b_r.

Standard gradient descent attempts to minimize (2.2) by updating the feed-forward weights following gradient directions according to

$$\beta_r(t+1) = \beta_r(t) - \eta \frac{\partial \mathcal{L}}{\partial \beta_r}(W(t), \beta(t))$$
$$w_r(t+1) = w_r(t) - \eta \frac{\partial \mathcal{L}}{\partial w_r}(W(t), \beta(t)),$$

for each $r \in [p]$, where $\eta > 0$ denotes the step size. We initialize $\beta(0)$ and $w_r(0)$ as standard Gaussian vectors. We introduce the notation $f(t), e(t) \in \mathbb{R}^n$, with $f_i(t) = f(x_i)$ denoting the network output on input x_i when the weights are $W(t)$ and $\beta(t)$, and $e_i(t) = y_i - f_i(t)$ denoting the corresponding prediction error or residual. With this notation, the gradients are expressed as

$$\frac{\partial \mathcal{L}}{\partial \beta_r} = \frac{1}{\sqrt{p}} \sum_{i=1}^n e_i \psi(w_r^\intercal x_i), \quad \frac{\partial \mathcal{L}}{\partial w_r} = \frac{1}{\sqrt{p}} \sum_{i=1}^n e_i \beta_r \psi'(w_r^\intercal x_i) x_i.$$

Here it is seen that the the gradient of the first-layer weights $\frac{\partial \mathcal{L}}{\partial w_r}$ involves not only the local input x_i and the change in the response of the r-th neuron, but also the backpropagated error signal $e_i \beta_r$. The appearance of β_r is, of course, due to the chain rule; but in effect it requires that the forward weights between layers are identical to the backward weights under error propagation. There is no evidence of biological mechanisms that would enable such "synaptic symmetry."

In the *feedback alignment* procedure of (Lillicrap et al., 2016), when updating the weights w_r, the error signal is weighted, and propagated backward, not by the second layer feedforward weights β, but rather by a random set of weights $b \in \mathbb{R}^p$ that are fixed during the course of training. Equivalently, the gradients for the first layer are replaced by the terms

$$\widetilde{\frac{\partial \mathcal{L}}{\partial w_r}} = \frac{1}{\sqrt{p}} \sum_{i=1}^n e_i b_r \psi'(w_r^\intercal x_i) x_i. \tag{2.3}$$

Note, however, that this update rule does not correspond to the gradient with respect to a modified loss function. The use of a random weight b_r when updating the first layer weights w_r does not violate locality, and could conceivably be implemented by biological mechanisms; we refer to Lillicrap et al. (2016); Bartunov et al. (2018); Lillicrap et al. (2020) for further discussion. A schematic of the relationship between the two algorithms is shown in Figure 1.

We can now summarize the main results and contributions of this paper. Our first result shows that the error converges to zero when using random backpropagation weights.

- Under Gaussian initialization of the parameters, if the model is sufficiently over-parameterized with $p \gg n$, then the error converges to zero linearly. Moreover, the parameters satisfy $\|w_r(t) - w_r(0)\| = \widetilde{O}\big(\frac{n}{\sqrt{p}}\big)$ and $|\beta_r(t) - \beta_r(0)| = \widetilde{O}\big(\frac{n}{\sqrt{p}}\big)$.

The precise assumptions and statement of this result are given in Theorem 3.2. The proof shows in the over-parameterized regime that the weights only change by a small amount. While related to

results for standard gradient descent, new methods are required because the "effective kernel" is not positive semi-definite.

We next turn to the issue of alignment of the second layer parameters β with the random backpropagation weights b. Such alignment was first observed in the original simulations of Lillicrap et al. (2016). With $h \in \mathbb{R}^p$ denoting the hidden layer of the two-layer network, the term $\delta_{\mathrm{BP}}(h) := \frac{\partial \mathcal{L}}{\partial h} = \frac{1}{\sqrt{p}} \beta \sum_{i=1}^n e_i$ represents how the error signals e_i are sent backward to update the feed-forward weights. With the use of random backpropagation weights, the error is instead propagated backward as $\delta_{\mathrm{FA}}(h) = \frac{1}{\sqrt{p}} b \sum_{i=1}^n e_i$.

Lillicrap et al. (2016) notice a decreasing angle between $\delta_{\mathrm{BP}}(h)$ and $\delta_{\mathrm{FA}}(h)$ during training, which is a sufficient condition to ensure that the algorithm converges. In the case of k-way classification, the last layer has k nodes, β and b are $p \times k$ matrices, and each error term e_i is a k-vector. In the regression setting, $k = 1$ so the angle between $\delta_{\mathrm{BP}}(h)$ and $\delta_{\mathrm{FA}}(h)$ is the same as the angle between β and b. Intuitively, the possibility for alignment is seen in the fact that while the updates for W use the error weighted by the random weights b, the updates for β indirectly involve W, allowing for the possibility that dependence on b will be introduced into β.

Our first result shows that, in fact, alignment will *not* occur in the over-parameterized setting. (So, while the error may still converge, "feedback alignment" may be a bit of a misnomer for the algorithm.)

- The cosine of the angle between the p-dimensional vectors δ_{FA} and δ_{BP} satisfies $\cos \angle(\delta_{\mathrm{FA}}, \delta_{\mathrm{BP}}(t)) = \cos \angle(b, \beta(t)) = O\big(\frac{n}{\sqrt{p}}\big)$.

However, we show that regularizing the parameters will cause δ_{BP} to align with δ_{FA} and therefore the parameters β to align with b. Since $\beta(0)$ and b are high dimensional Gaussian vectors, they are nearly orthogonal with high probability. The effect of regularization can be seen as shrinking the component of $\beta(0)$ in the parameters over time. Our next result establishes this precisely in the linear case.

- Supposing that $\psi(u) = u$, then introducing a ridge penalty $\lambda(t)\|\beta\|^2$ where $\lambda(t) = \lambda$ for $t \le T$ and $\lambda(t) = 0$ for $t > T$ on β causes the parameters to align, with $\cos \angle(b, \beta(t)) \ge c > 0$ for sufficiently large t.

The technical conditions are given in Theorem 4.6. Our simulations are consistent with this result, and also show alignment with a constant regularization $\lambda(t) \equiv \lambda$, for both linear and nonlinear activation functions. Finally, we complement this result by showing that convergence is preserved with regularization, for general activation functions. This is presented in Theorem 4.2.

3 Convergence with Random Backpropagation Weights

Due to the replacement of backward weights with the random backpropagation weights, there is no guarantee *a priori* that the algorithm will reduce the squared error loss $\mathcal{L}$. Lillicrap et al. (2020) study the convergence on two-layer linear networks in a continuous time setting. Through the analysis of a system of differential equations on the network parameters, convergence to the true linear target function is shown, in the population setting of arbitrarily large training data. Among recent studies of over-parametrized networks under backpropagation, the neural tangent kernel (NTK) is heavily utilized to describe the evolution of the network during training (Jacot et al., 2018; Chen & Xu, 2020). For any neural network $f(x, \theta)$ with parameter θ, the NTK is defined as

$$K_f(x, y) = \Big\langle \frac{\partial f(x, \theta)}{\partial \theta}, \frac{\partial f(y, \theta)}{\partial \theta} \Big\rangle.$$

Given a dataset $\{(x_i, y_i)\}_{i=1}^n$, we can also consider its corresponding Gram matrix $K = (K_f(x_i, x_j))_{n \times n}$. Jacot et al. (2018) show that in the infinite width limit, K_f converges to a constant at initialization and does not drift away from initialization throughout training. In the over-parameterized setting, if the Gram matrix K is positive definite, then K will remain close to its initialization during training, resulting in linear convergence of the squared error loss (Du et al., 2018, 2019; Gao & Lafferty, 2020). For the two-layer network $f(x, \theta)$ defined in (2.1) with $\theta = (\beta, W)$,

the kernel K_f can be written in two parts, G_f and H_f, which correspond to β and W respectively:

$$K_f(x,y) = G_f(x,y) + H_f(x,y) := \left\langle \frac{\partial f(x,\theta)}{\partial \beta}, \frac{\partial f(y,\theta)}{\partial \beta} \right\rangle + \sum_{r=1}^{p} \left\langle \frac{\partial f(x,\theta)}{\partial w_r}, \frac{\partial f(y,\theta)}{\partial w_r} \right\rangle.$$

Under the feedback alignment scheme with random backward weights b, G_f remains the same as for standard backpropagation, while one of the gradient terms $\frac{\partial f}{\partial w_r}$ in H_f changes to $\widetilde{\frac{\partial f(x,\theta)}{\partial w_r}} = \frac{1}{\sqrt{p}} b_r \psi'(w_r^\intercal x)x$, with H_f replaced by $H_f = \sum_{r=1}^{p} \left\langle \widetilde{\frac{\partial f(x,\theta)}{\partial w_r}}, \frac{\partial f(y,\theta)}{\partial w_r} \right\rangle$. As a result, H_f is no longer positive semi-definite and close to 0 at initialization if the network is over-parameterized. However, if $G = (G_f(x_i, x_j))_{n\times n}$ is positive definite and $H = (H_f(x_i, x_j))_{n\times n}$ remains small during training, we are still able to show that the loss $\mathcal{L}$ will converge to zero exponentially fast.

Assumption 3.1. Define the matrix $\overline{G} \in \mathbb{R}^{n\times n}$ with entries $\overline{G}_{i,j} = \mathbb{E}_{w\sim\mathcal{N}(0,I_p)} \psi(w^\intercal x_i)\psi(w^\intercal x_j)$. Then we assume that the minimum eigenvalue satisfies $\lambda_{\min}(\overline{G}) \geq \gamma$, where γ is a positive constant.

Theorem 3.2. *Let $W(0)$, $\beta(0)$ and b have i.i.d. standard Gaussian entries. Assume* (1) *Assumption 3.1 holds,* (2) *ψ is smooth, ψ, ψ' and ψ'' are bounded and* (3) *$|y_i|$ and $\|x_i\|$ are bounded for all $i \in [n]$. Then there exists positive constants c_1, c_2, C_1 and C_2, such that for any $\delta \in (0,1)$, if $p \geq \max\left(C_1 \frac{n^2}{\delta\gamma^2}, C_2 \frac{n^4 \log p}{\gamma^4}\right)$, then with probability at least $1-\delta$ we have that*

$$\|e(t+1)\| \leq (1 - \frac{\eta\gamma}{4})\|e(t)\| \tag{3.1}$$

and

$$\|w_r(t) - w_r(0)\| \leq c_1 \frac{n\sqrt{\log p}}{\gamma\sqrt{p}}, \quad |\beta_r(t) - \beta_r(0)| \leq c_2 \frac{n}{\gamma\sqrt{p}} \tag{3.2}$$

for all $r \in [p]$ and $t > 0$.

We note that the matrix $\overline{G}$ in Assumption 3.1 is the expectation of G with respect to the random initialization, and is thus close to $\overline{G}$ due to concentration. To justify the assumption, we provide the following proposition, which states that Assumption 3.1 holds when the inputs x_i are drawn independently from a Gaussian distribution. The proofs of Theorem 3.2 and Proposition 3.3 are deferred to Appendix A.

Proposition 3.3. *Suppose $x_1, ..., x_n \overset{i.i.d.}{\sim} \mathcal{N}(0, I_d/d)$ and the activation function ψ is sigmoid or tanh. If $d = \Omega(n)$, then Assumption 3.1 holds with high probability.*

4 Alignment with Random Backpropagation Weights

The most prominent characteristic of the feedback alignment algorithm is the phenomenon that the error signals propagated with the forward weights align with those propagated with fixed random backward weights during training. Specifically, if we denote $h \in \mathbb{R}^p$ to be the hidden layer of the network, then we write $\delta_{\text{BP}}(h) := \frac{\partial \mathcal{L}}{\partial h}$ to represent the error signals with respect to the hidden layer that are backpropagated with the feed-forward weights and $\delta_{\text{FA}}(h)$ as the error signals computed with fixed random backward weights. In particular, the error signals $\delta_{\text{BP}}(h)$ and $\delta_{\text{FA}}(h)$ for the two-layer network (2.1) are given by

$$\delta_{\text{BP}}(h) = \frac{1}{\sqrt{p}} \beta \sum_{i=1}^{n} e_i \quad \text{and} \quad \delta_{\text{FA}}(h) = \frac{1}{\sqrt{p}} b \sum_{i=1}^{n} e_i.$$

Lillicrap et al. (2016) notice a decreasing angle between $\delta_{\text{BP}}(h)$ and $\delta_{\text{FA}}(h)$ during training. We formalize this concept of alignment by the following definition.

Definition 4.1. We say a two-layer network *aligns* with the random weights b during training if there exists a constant $c > 0$ and time T_c such that $\cos\angle(\delta_{\text{FA}}, \delta_{\text{BP}}(t)) = \cos\angle(b, \beta(t)) = \frac{\langle b, \beta(t)\rangle}{\|b\|\|\beta(t)\|} \geq c$ for all $t > T_c$.

4.1 Regularized feedback alignment

Unfortunately, alignment between $\beta(t)$ and b is not guaranteed for over-parameterized networks and the loss (2.2). In particular, we control the cosine value of the angle by inequalities (3.2) from Theorem 3.2, *i.e.*,

$$\left|\cos\angle(b,\beta(t))\right| \le \frac{|\langle \frac{b}{\|b\|},\beta(0)\rangle| + \|\beta(t)-\beta(0)\|}{\|\beta(0)\| - \|\beta(t)-\beta(0)\|} = O\left(\frac{n}{\sqrt{p}}\right),$$

which indicates that $\beta(t)$ and b become orthogonal as the network becomes wider. Intuitively, this can be understood as resulting from the parameters staying near their initializations during training when p is large, where $\beta(0)$ and b are almost orthogonal to each other. This motivates us to regularize the network parameters. We consider in this work the squared error loss with an ℓ_2 regularization term on β:

$$\mathcal{L}(t,W,\beta) = \frac{1}{2}\sum_{i=1}^{n}\big(f(x_i)-y_i\big)^2 + \frac{1}{2}\lambda(t)\|\beta\|^2, \tag{4.1}$$

where $\{\lambda(t)\}_{t=0}^{\infty}$ is a sequence of regularization rates, which defines a series of loss functions for different training steps t. Thus, the update for w_r remains the same and the update for β changes to

$$\beta_r(t+1) = (1-\lambda(t))\beta_r(t) - \frac{\eta}{\sqrt{p}}\sum_{i=1}^{n} e_i(t)\psi(w_r(t)^\intercal x_i), \quad \text{for } r\in[p].$$

Comparing to Algorithm 1, an extra contraction factor $1-\lambda(t)$ is added in the update of $\beta(t)$, which doesn't affect the locality of the algorithm but helps the alignment by shrinking the component of $\beta(0)$ in $\beta(t)$.

Following Theorem 3.2, we provide an error bound for regularized feedback alignment in Theorem 4.2. Since regularization terms $\lambda(t)$ make additional contributions to the error $e(t)$ as well as to the kernel matrix G, an upper bound on $\sum_{t\ge0}\lambda(t)$ is needed to ensure positivity of the minimal eigenvalue of G during training, in order for the error $e(t)$ to be controlled. In particular, if there is no regularization, *i.e.*, $\lambda(t)=0$ for all $t\ge0$, then we recover exponential convergence for the error $e(t)$ as in Theorem 3.2. The proof of Theorem 4.2 is also deferred to Appendix A.

Theorem 4.2. *Assume all the conditions from Theorem 3.2. Assume* $\sum_{t=0}^{\infty}\lambda(t) \le \tilde{S}_\lambda = \tilde{c}_S\frac{\gamma^2\sqrt{p}}{\eta n^2\sqrt{\log p}}$ *for some constant* $\tilde{c}_S$. *Then there exist positive constants* C_1 *and* C_2, *such that for any* $\delta\in(0,1)$, *if* $p\ge\max\big(C_1\frac{n^2}{\delta\gamma^2}, C_2\frac{n^4\log p}{\gamma^4}\big)$, *then with probability at least* $1-\delta$, *we have*

$$\|e(t+1)\| \le \Big(1-\frac{\eta\gamma}{4}-\eta\lambda(t)\Big)\|e(t)\| + \lambda(t)\|y\| \tag{4.2}$$

for all $t\ge0$.

4.2 Alignment analysis for linear networks

In this section, we focus on the theoretical analysis of alignment for linear networks, which is equivalent to setting the activation function ψ to the identity map. The loss function can be written as

$$\mathcal{L}(t,W,\beta) = \frac{1}{2}\Big\|\frac{1}{\sqrt{p}}XW^\intercal\beta - y\Big\|^2 + \frac{\lambda(t)}{2}\|\beta\|^2,$$

where $X=(x_1,\dots,x_n)^\intercal$; this is a form of over-parameterized ridge regression. Before presenting our results on alignment, we first provide a linear version of Theorem 4.2 that adopts slightly different conditions.

Theorem 4.3. *Assume* (1) $\|y\|=\Theta(\sqrt{n})$, $\lambda_{\min}(XX^\intercal)>\gamma$ *and* $\lambda_{\max}(XX^\intercal)<M$ *for some constants* $M>\gamma>0$, *and* (2) $\sum_{t=0}^{\infty}\lambda(t)\le S_\lambda = c_S\frac{\gamma\sqrt{\gamma p}}{\eta\sqrt{n}M}$ *for some constant* c_S. *Then for any* $\delta\in(0,1)$, *if* $p=\Omega(\frac{Md\log(d/\delta)}{\gamma})$, *the following inequality holds for all* $t\ge0$ *with probability at least* $1-\delta$*:*

$$\|e(t+1)\| \le \big(1-\frac{\eta\gamma}{2}-\eta\lambda(t)\big)\|e(t)\| + \lambda(t)\|y\|. \tag{4.3}$$

We remark that in the linear case, the kernel matrix G reduces to the form $XW^\intercal WX^\intercal$ and its expectation $\overline{G}$ at initialization also reduces to $XX^\intercal$. Thus, Assumption 3.1 holds if $XX^\intercal$ is positive definite, which is equivalent to the x_i's being linearly independent. The result of Theorem 4.2 can not be directly applied to the linear case since we assume that ψ is bounded, which is true for sigmoid or $\tanh$ but not for the identity map. This results in a slightly different order for S_λ and an improved order for p.

Our results on alignment also rely on an isometric condition on X, which requires the minimum and the maximum eigenvalues of $XX^\intercal$ to be sufficiently close (*cf.* Definition 4.4). On the other hand, this condition is relatively mild and can be satisfied when X has random Gaussian entries with a gentle dimensional constraint, as demonstrated by Proposition 4.5. Finally, we show in Theorem 4.6 that under a simple regularization strategy where a constant regularization is adopted until a cutoff time T, regularized feedback alignment achieves alignment if X satisfies the isometric condition.

Definition 4.4 ((γ, ε)-Isometry)**.** Given positive constants γ and ε, we say X is (γ, ε)-isometric if $\lambda_{\min}(XX^\intercal) \geq \gamma$ and $\lambda_{\max}(XX^\intercal) \leq (1+\varepsilon)\gamma$.

Proposition 4.5. *Assume $X \in \mathbb{R}^{n\times d}$ has independent entries drawn from $N(0, 1/d)$. For any $\varepsilon \in (0, 1/2)$ and $\delta \in (0, 1)$, if $d = \Omega(\frac{1}{\varepsilon}\log\frac{n}{\delta} + \frac{n}{\varepsilon}\log\frac{1}{\varepsilon})$, then X is $(1-\varepsilon, 4\varepsilon)$-isometric with probability $1-\delta$.*

Theorem 4.6. *Assume all conditions from Theorem 4.3 hold and X is (γ, ε)-isometric with a small constant ε. Let the regularization weights satisfy*

$$\lambda(t) = \begin{cases} \lambda, & t \leq T, \\ 0, & t > T, \end{cases}$$

with $\lambda = L\gamma$ and $T = \lfloor S_\lambda/\lambda \rfloor$ for some large constant L. Then for any $\delta \in (0, 1)$, if $p = \Omega(d\log(d/\delta))$, with probability at least $1-\delta$, regularized feedback alignment achieves alignment. Specifically, there exist a positive constant $c = c_\delta$ and time T_c, such that $\cos\angle(b, \beta(t)) \geq c$ for all $t > T_c$.

We defer the proofs of Proposition 4.5, Theorem 4.3 and Theorem 4.6 to Appendix B. In fact, we prove Theorem 4.6 by directly computing $\beta(t)$ and the cosine of the angle. Although b doesn't show up in the update of β, it can still propagate to β through W. Since the size of the component of b in $\beta(t)$ depends on the inner-product $\langle e(t), e(t')\rangle$ for all previous steps $t' \leq t$, the norm bound (4.3) from Theorem 4.3 is insufficient; thus, a more careful analysis of $e(t)$ is required.

We should point out that the constant c in the lower bound is independent of the sample size n, input dimension d, network width p and learning rate η. We also remark that the cutoff schedule of $\lambda(t)$ is just chosen for simplicity. For other schedules such as inverse-squared decay or exponential decay, one could also obtain the same alignment result as long as the summation of $\lambda(t)$ is less than S_λ.

Large sample scenario. In Theorems 4.3 and 4.6, we consider the case where the sample size n is less than the input dimension d, so that positive definiteness of $XX^\intercal$ can be established. However, both results still hold for $n > d$. In fact, the squared error loss $\mathcal{L}$ can be written as

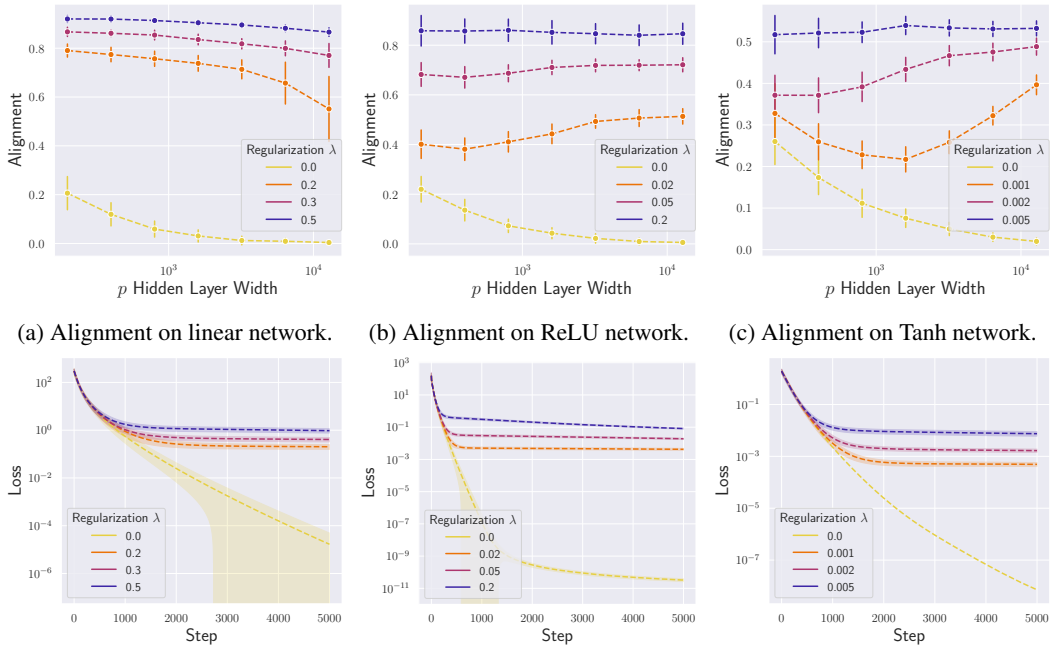
$$\sum_{i=1}^{n} \big(f(x_i) - y\big)^2 = \big\|\frac{1}{\sqrt{p}} XW^\intercal\beta - y\big\|^2 = \big\|\frac{1}{\sqrt{p}} XW^\intercal\beta - \bar{y}\big\|^2 + \|\bar{y} - y\|^2,$$

where $\bar{y}$ denotes the projection of y onto the column space of X. Without loss of generality, we assume $y = \bar{y}$. As a result, y and the columns of X are all in the same d-dimensional subspace of $\mathbb{R}^n$ and $XX^\intercal$ is positive definite on this subspace, as long as X has full column rank. Consequently, we can either work on this subspace of $\mathbb{R}^n$ or project all the vectors onto $\mathbb{R}^d$, and the isometric condition is revised to only consider the d nonzero eigenvalues of $XX^\intercal$.

5 Simulations

Our experiments apply the feedback alignment algorithm to two-layer networks, using a range of networks with different widths and activations. The numerical results suggest that regularization is essential in achieving alignment, in both regression and classification tasks, for linear and nonlinear models. We implement the feedback alignment procedure in PyTorch as an extension of the autograd module for backpropagation, and the training is done on V100 GPUs from internal clusters.

Feedback alignment on synthetic data. We first train two-layer networks on synthetic data, where each network f shares the architecture shown in (2.1) and the data are generated by another network f_0 that has the same architecture but with random Gaussian weights. We present the experiments for both linear and nonlinear networks, where the activation functions are chosen to be Rectified Linear Unit (ReLU) and hyperbolic tangent (Tanh) for nonlinear case. We set training sample sample size to $n = 50$ and the input dimension $d = 150$, but vary the hidden layer width $p = 100 \times 2^k$ with $k \in [7]$. During training, we take step size $\eta = 10^{-4}$ for linear networks and $\eta = 10^{-3}, 10^{-2}$ for ReLU and Tanh networks, respectively.



(a) Alignment on linear network. (b) Alignment on ReLU network. (c) Alignment on Tanh network.

(d) Loss on linear network. (e) Loss on ReLU network. (f) Loss on Tanh network.

Figure 2: Comparisons of alignment and convergence for the feedback alignment algorithm with different levels of ℓ_2 regularization. In Figs. 2a to 2c, the data points represent the mean value computed across simulations, and the error bars mark the standard deviation out of 50 independent runs. In Figs. 2d to 2f, we show the trajectories of the training loss for networks with $p = 3200$, with the shaded areas indicating the standard deviation over 50 independent runs. The x-axes on the first row and the y-axes on the second row are presented using a logarithmic scale.

In Figs. 2a to 2c, we show how alignment depends on regularization and the degree of overparameterization as measured by the hidden layer width p. Alignment is measured by the cosine of the angle between the forward weights β and backward weights b. We train the networks until the loss function converges; this procedure is repeated 50 times for each p and λ. For all three types of networks, as p increases, alignment vanishes if there is no regularization, and grows with the level of regularization λ for the same network. We complement the alignment plots with the corresponding loss curves, where the training loss converges slower with larger regularization. These numerical results are consistent with our theoretical statements. Due to the regularization, the loss converges to a positive number that is of the same order as λ.

We remark that using dropout as a form of regularization can also help the alignment between forward and backward weights (Wager et al., 2013). However, our numerical results suggest that dropout regularization fails to keep the alignment away from zero for networks with large hidden layer width. No theoretical result is available that explains the underlying mechanism.

Feedback alignment on the MNIST dataset. The `MNIST` dataset is available under the Creative Commons Attribution-Share Alike 3.0 license (Deng, 2012). It consists of 60,000 training images and 10,000 test images of dimension 28 by 28. We reshape them into vectors of length $d = 784$ and

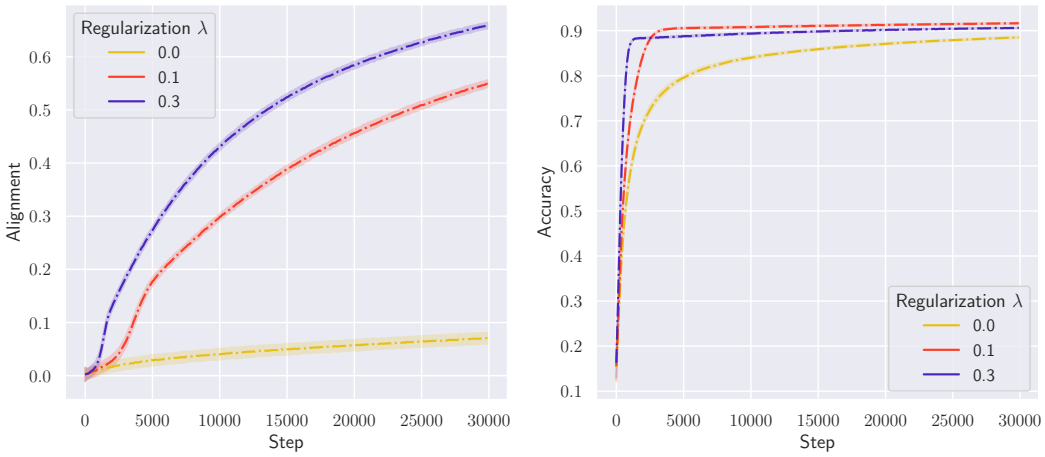


Figure 3: Comparisons on alignment and accuracy for feedback alignment algorithm with $\lambda = 0, 0.1, 0.3$. The left figure shows alignment defined by $\cos\angle(\delta_{\mathrm{BP}}(h), \delta_{\mathrm{FA}}(h))$, and right figure shows the accuracy on the test set. The dashed lines and corresponding shaded areas represent the means and the standard deviations over 10 runs with random initialization.

normalize them by their mean and standard deviation. The network structure is 784-1000-10 with ReLU activation at the hidden layer and with softmax normalization at output layer. During training, we choose the batch size to be 600 and the step size $\eta = 10^{-2}$. The training procedure uses 300 epochs in total. We repeat the training 10 times for each choice of λ.

Fig. 3 shows the performance of feedback alignment with regularization $\lambda = 0, 0.1, 0.3$. Since the output of the network is not one-dimensional but 10-dimensional, the alignment is now measured by $\cos\angle(\delta_{\mathrm{BP}}(h), \delta_{\mathrm{FA}}(h))$, where $\delta_{\mathrm{BP}}(h)$ is the error signal propagated to the hidden neurons h through forward weights β, and $\delta_{\mathrm{FA}}(h)$ the error weighted by the random backward weights b. We observe that both alignment and convergence are improved by adding regularization to the training, and increasing the regularization level λ can further facilitate alignment, with a small gain in test accuracy.

6 Discussion

In this paper we have analyzed the feedback alignment algorithm of Lillicrap et al. (2016), showing convergence of the algorithm. The convergence is subtle, as the algorithm does not directly minimize the target loss function; rather, the error is transferred to the hidden neurons through random weights that do not change during the course of learning. The supplement to Lillicrap et al. (2016) presents interesting insights on the dynamics of the algorithm, such as how the feedback weights act as pseudoinverse of the forward weights. After giving an analysis of convergence in the linear case, the authors state that "a general proof must be radically different from those used to demonstrate convergence for backprop" (Supplementary note 16), observing that the algorithm does not minimize any loss function. Our proof of convergence in the general nonlinear case leverages techniques from the use of neural tangent kernel analysis in the over-parameterized setting, but requires more care because the kernel is not positive semi-definite at initialization. In particular, as a sum of two terms G and H, the matrix G is concentrated around its postive-definite mean, while H is not generally postive-semidefinite. However, we show that the entries of both matrices remain close to their initial values, due to over-parameterization, and analyze the error term in a Taylor expansion, which establishes convergence.

In analyzing alignment, we unexpectedly found that regularization is essential; without it, the alignment may not persist as the network becomes wider, as our simulations clearly show. Our analysis in the linear case proceeds by essentially showing that

$$\beta(t) = (1 - \eta\lambda)^{t-1}\beta(0) + \frac{\eta}{\sqrt{p}} W(0) X^{\intercal} \alpha_1(t-1) + \left(\frac{\eta}{\sqrt{p}}\right) b\alpha_2(t-1)$$

and controlling α_1 while showing that α_2 remains sufficiently large; the regularization kills off the first term. Although we see no obstacle, in principle, to carrying out this proof strategy in the nonlinear case, the calculations are more complex. While convergence requires analysis of the norm of the error, alignment requires understanding the direction of the error. But our simulations strongly suggest this result will go through.

In terms of future research, a technical direction is to extend our results to multilayer networks. It would be interesting to explore local methods to update the backward weights b, rather than fixing them, perhaps using a Hebbian update rule in combination with the forward weights W. More generally, it is important to study other biologically plausible learning rules that can be implemented in deep learning frameworks at scale and without loss of performance. The results presented here offer support for this as a fruitful line of research. Biologically plausible computational learning contributes to, and shares societal impact with, a large body of fundamental research that aims to understand the basis for cognition in animals, including humans.

Acknowledgments

Research supported in part by NSF grant CCF-1839308.

References

Akrout, M., Wilson, C., Humphreys, P., Lillicrap, T., & Tweed, D. B. (2019). Deep learning without weight transport. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, volume 32: Curran Associates, Inc.

Bartunov, S., Santoro, A., Richards, B., Marris, L., Hinton, G. E., & Lillicrap, T. (2018). Assessing the scalability of biologically-motivated deep learning algorithms and architectures. In *Advances in Neural Information Processing Systems* (pp. 9368–9378).

Bellec, G., Scherr, F., Hajek, E., Salaj, D., Legenstein, R., & Maass, W. (2019). Biologically inspired alternatives to backpropagation through time for learning in recurrent neural nets.

Chen, L. & Xu, S. (2020). Deep neural tangent kernel and laplace kernel have the same rkhs. *arXiv preprint arXiv:2009.10683*.

Deng, L. (2012). The MNIST database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6), 141–142.

Du, S., Lee, J., Li, H., Wang, L., & Zhai, X. (2019). Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning* (pp. 1675–1685).: PMLR.

Du, S. S., Zhai, X., Poczos, B., & Singh, A. (2018). Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*.

Elman, J. L., Bates, E. A., Johnson, M. H., Annette Karmiloff-Smith, D. P., & Plunkett, K. (1996). *Rethinking Innateness: A connectionist perspective on development*. Cambridge MA: MIT Press.

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

Gao, C. & Lafferty, J. (2020). Model repair: Robust recovery of over-parameterized statistical models. *arXiv preprint arXiv:2005.09912*.

Hand, P. & Voroninski, V. (2018). Global guarantees for enforcing deep generative priors by empirical risk. In *Conference On Learning Theory* (pp. 970–978).: PMLR.

Hebb, D. O. (1961). Distinctive features of learning in the higher animal. In J. F. Delafresnaye (Ed.), *Brain Mechanisms and Learning*: London: Oxford University Press.

Jacot, A., Gabriel, F., & Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572*.

Launay, J., Poli, I., Boniface, F., & Krzakala, F. (2020). Direct feedback alignment scales to modern deep learning tasks and architectures. *arXiv preprint arXiv:2006.12878*.

Laurent, B. & Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, (pp. 1302–1338).

Lillicrap, T. P., Cownden, D., Tweed, D. B., & Akerman, C. J. (2016). Random synaptic feedback weights support error backpropagation for deep learning. *Nature communications*, 7(1), 1–10.

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

Medler, D. A. (1998). A brief history of connectionism. *Neural Computing Surveys*, 1, 61–101.

Nøkland, A. (2016). Direct feedback alignment provides learning in deep neural networks. *arXiv preprint arXiv:1609.01596*.

Oja, E. (1982). A simplified neuron model as a principal component analyzer. *J. Mathematical Biology*, 15, 267–273.

Paulsen, O. & Sejnowski, T. J. (2000). Natural patterns of activity and long-term synaptic plasticity. *Current Opinion in Neurobiology*, 10(2), 172–179.

Rumelhart, D., McClelland, J., & the PDP Research Group (1986a). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 2: Psychologcal and Biological Models. Cambridge, Massachusetts: MIT Press.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986b). Learning representations by back-propagating errors. *Nature*, 323(6088), 533–536.

Sejnowski, T. J. (1999). The book of Hebb. *Neuron*, 24, 773–776.

Sejnowski, T. J. & Tesauro, G. (1989). The hebb rule for synaptic plasticity: Algorithms and implementations. In J. H. Byrne & W. O. Berry (Eds.), *Neural Models of Plasticity* (pp. 94–103).

Wager, S., Wang, S., & Liang, P. (2013). Dropout training as adaptive regularization. *arXiv preprint arXiv:1307.1493*.

Yamins, D. & DiCarlo, J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3), 356–365.

Yildirim, I., Wu, J., Kanwisher, N., & Tenenbaum, J. (2019). An integrative computational architecture for object-driven cortex. *J.B. Current Opinion in Neurobiology*.

A Convergence on Two-Layer Nonlinear Networks

We consider the family of neural networks

$$f(x) = \frac{1}{\sqrt{p}} \sum_{r=1}^{p} \beta_r \psi(w_r^\intercal x) = \frac{1}{\sqrt{p}} \beta^\intercal \psi(Wx) \tag{A.1}$$

where $\beta \in \mathbb{R}^p$, $W = (w_1, ..., w_p)^\intercal \in \mathbb{R}^{p \times d}$, and ψ is an activation function. Given data, the loss function is

$$\mathcal{L}(W, \beta) = \frac{1}{2} \sum_{i=1}^{n} (f(x_i) - y_i)^2 = \frac{1}{2} \sum_{i=1}^{n} \Big(\frac{1}{\sqrt{p}} \beta^\intercal \psi(W x_i) - y \Big)^2. \tag{A.2}$$

The feedback alignment algorithm has updates

$$\begin{aligned} W(t+1) &= W(t) - \eta \frac{1}{\sqrt{p}} \sum_{i=1}^{n} D_i(t) b x_i^\intercal e_i(t) \\ \beta(t+1) &= \beta(t) - \eta \frac{1}{\sqrt{p}} \sum_{i=1}^{n} \psi(W(t) x_i) e_i(t) \end{aligned} \tag{A.3}$$

where $D_i(t) = \text{diag}(\psi'(W(t)x_i))$ and $e_i(t) = \frac{1}{\sqrt{p}} \beta(t)^\intercal \psi(W(t)x_i) - y_i$. To help make the proof more readable, we use c, C to denote the global constants whose values may vary from line to line.

A.1 Concentration Results

Lemma A.1 (Lemma A.7 in Gao & Lafferty, 2020)**.** *Assume* $x_1, ..., x_n \overset{i.i.d.}{\sim} \mathcal{N}(0, I_d/d)$. *We define matrix* $\widetilde{G} \in \mathbb{R}^{n \times n}$ *with entries*

$$\widetilde{G}_{i,j} = |\mathbb{E}\psi'(Z)|^2 \frac{x_i^\intercal x_j}{\|x_i\| \|x_j\|} + (\mathbb{E}|\psi(Z)|^2 - |\mathbb{E}\psi'(Z)|^2) \mathbb{I}\{i = j\}$$

where $Z \sim \mathcal{N}(0, 1)$. *If* $d = \Omega(\log n)$, *then with high probability, we have*

$$\|\overline{G} - \widetilde{G}\|^2 \lesssim \frac{\log n}{d} + \frac{n^2}{d^2}.$$

Proof of Proposition 3.3. If ψ is sigmoid or tanh, for a standard Gaussian random variable Z, we have

$$\gamma := \frac{1}{2}(\mathbb{E}|\psi(Z)|^2 - |\mathbb{E}\psi'(Z)|^2) > 0.$$

From Lemma A.1, we know that with high probability $\lambda_{\min}(\overline{G}) \geq \lambda_{\min}(\widetilde{G}) - \|\overline{G} - \widetilde{G}\| \geq 2\gamma - C(\sqrt{\frac{\log n}{d}} + \frac{n}{d}) \geq \gamma$. □

Lemma A.2. *Assume* $W(0)$, $\beta(0)$ *and* b *have i.i.d. standard Gaussian entries. Given* $\delta \in (0, 1)$, *if* $p = \Omega(n/\delta)$, *then with probability* $1 - \delta$

$$\frac{1}{p} \sum_{r=1}^{p} |b_r| \leq c, \tag{A.4}$$

$$\frac{1}{p} \sum_{r=1}^{p} |b_r \beta_r(0)| \leq c, \tag{A.5}$$

$$\|e(0)\| \leq c\sqrt{n}, \tag{A.6}$$

$$\max_{r \in [p]} |b_r| \leq 2\sqrt{\log p}. \tag{A.7}$$

Proof. We will show each inequality holds with probability at least $1-\frac{\delta}{4}$, then by a union bound, all of them hold with probability at least $1-\delta$. Since $\mathbb{V}\mathrm{ar}(\frac{1}{p}\sum_{r=1}^{p}|b_r|) \le \frac{\mathbb{V}\mathrm{ar}(|b_0|)}{p}$, by Chebyshev's inequality, we have

$$\mathbb{P}(\frac{1}{p}\sum_{r=1}^{p}|b_r| > \mathbb{E}(b_1)+1) \le \frac{\mathbb{V}\mathrm{ar}(|b_1|)}{p} \le \delta/4$$

if $p \ge 4\mathbb{V}\mathrm{ar}(|b_1|)/\delta$, which gives (A.4). The proof for (A.5) is similar since $\mathbb{V}\mathrm{ar}(\frac{1}{p}\sum_{r=1}^{p}|b_r\beta_r(0)|) = O(1/p)$. To prove (A.6), since $|y_i|$ and $\|x_i\|$ are bounded, it suffices to show $|u_i(0)| \le c$ for all $i \in [n]$. Actually, by independence, we have

$$\mathbb{V}\mathrm{ar}(u_i(0)) = \mathbb{V}\mathrm{ar}\Big(\frac{1}{p}\sum_{r=1}^{p}\beta_r(0)\psi(w_r(0)^\intercal x_i)\Big) = \frac{1}{p}\mathbb{V}\mathrm{ar}\Big(\beta_1(0)\psi(w_1(0)^\intercal x_i)\Big) = O(1/p).$$

By Chebyshev's inequality, we have for each $i \in [n]$

$$\mathbb{P}(|u_i(0)| > c) \le \frac{\mathbb{V}\mathrm{ar}(u_i(0))}{c^2} \le \frac{\delta}{4n}$$

where we require $p = \Omega(n/\delta)$. With a union bound argument, we can show (A.6). Finally, (A.7) followed from standard Gaussian tail bounds and union bound argument, yielding

$$\mathbb{P}(\max_{r\in[p]}|b_r| > 2\sqrt{\log p}) \le \sum_{r\in[p]}\mathbb{P}(|b_r| > 2\sqrt{\log p}) \le 2pe^{-2\log p} = \frac{2}{p} \le \frac{\delta}{4}.$$

□

Lemma A.3. *Under the conditions of Theorem 3.2, we define matrices $G(0), H(0) \in \mathbb{R}^{n\times n}$ with entries*

$$G_{ij}(0) = \frac{1}{p}\psi(W(0)x_i)^\intercal\psi(W(0)x_j) = \frac{1}{p}\sum_{r=1}^{p}\psi(w_r(0)^\intercal x_i)\psi(w_r(0)^\intercal x_j) \tag{A.8}$$

and

$$H_{ij}(0) = \frac{x_i^\intercal x_j}{p}\beta(0)^\intercal D_i(0)D_j(0)b = \frac{1}{p}\sum_{r=1}^{p}\beta_r(0)b_r\psi'(w_r(0)^\intercal x_i)\psi'(w_r(0)^\intercal x_j). \tag{A.9}$$

For any $\delta \in (0,1)$, if $p = \Omega(\frac{n^2}{\delta\gamma^2})$, then with probability at least $1-\delta$, we have $\lambda_{\min}(G(0)) \ge \frac{3}{4}\gamma$ and $\|H(0)\| \le \frac{\gamma}{4}$.

Proof. By independence and boundedness of ψ and ψ', we have $\mathbb{V}\mathrm{ar}(G_{ij}(0)) = O(1/p)$ and $\mathbb{V}\mathrm{ar}(H_{ij}(0)) = O(1/p)$. Since $\mathbb{E}(G(0)) = \overline{G}$, we have

$$\mathbb{E}\|G(0)-\overline{G}\|^2 \le \mathbb{E}\|G(0)-\overline{G}\|_F^2 = O(\frac{n^2}{p}).$$

By Markov's inequality, when $p = \Omega(\frac{n^2}{\delta\gamma^2})$

$$\mathbb{P}(\|G(0)-\overline{G}\| > \frac{\gamma}{4}) \le O(\frac{n^2}{p\gamma^2}) \le \frac{\delta}{2}.$$

Similarly we have $\mathbb{P}(\|H(0)\| > \frac{\gamma}{4}) \le \frac{\delta}{2}$, since $\mathbb{E}(H(0)) = 0$. Then with probability at least $1-\delta$, $\lambda_{\min}(G(0)) \ge \lambda_{\min}(\overline{G}) - \gamma/4 \ge \frac{3}{4}\gamma$, and $\|H(0)\| \le \gamma/4$. □

A.2 Proof of Theorem 3.2

Lemma A.4. *Assume all the inequalities from Lemma A.2 hold. Under the conditions of Theorem 3.2, if the error bound (3.1) holds for all $t = 1, 2, ..., t'-1$, then the bounds (3.2) hold for all $t \le t'$.*

Proof. From the feedback alignment updates (A.3), we have for all $t \leq T$

$$\begin{aligned}
|\beta_r(t) - \beta_r(0)| &\leq \frac{\eta}{\sqrt{p}} \sum_{s=0}^{t-1} \sum_{i=1}^{n} |\psi(w_r(t)x_i)e_i(t)| \\
&\leq c\frac{\eta}{\sqrt{p}} \sum_{s=0}^{t-1} \sum_{i=1}^{n} |e_i(t)| \\
&\leq c\frac{\eta\sqrt{n}}{\sqrt{p}} \sum_{s=0}^{t-1} \|e(t)\| \\
&\leq c\frac{\eta\sqrt{n}}{\sqrt{p}} \sum_{s=0}^{t-1} (1 - \frac{\gamma\eta}{4})^t \|e(0)\| \\
&\leq c\frac{\sqrt{n}}{\gamma\sqrt{p}} \|e(0)\| \\
&\leq c\frac{n}{\gamma\sqrt{p}}
\end{aligned}$$

where we use the fact that ψ is bounded and (A.6). We also have

$$\begin{aligned}
\|w_r(t) - w_r(0)\| &\leq \frac{\eta}{\sqrt{p}} \sum_{s=0}^{t-1} \sum_{i=1}^{n} \|\psi'(w_r(t)^\intercal x_i)b_r x_i e_i(t)\| \\
&\leq c\frac{\eta}{\sqrt{p}} \sum_{s=0}^{t-1} \sum_{i=1}^{n} |b_r||e_i(t)| \\
&\leq c|b_r|\frac{\eta\sqrt{n}}{\sqrt{p}} \sum_{s=0}^{t-1} \|e(t)\| \\
&\leq c|b_r|\frac{\sqrt{n}}{\gamma\sqrt{p}} \|e(0)\| \\
&\leq c\frac{n\sqrt{\log p}}{\gamma\sqrt{p}}
\end{aligned}$$

where we use that ψ' is bounded, (A.6) and (A.7). □

Lemma A.5. *Assume all the inequalities from Lemma A.2 hold. Under the conditions of Theorem 3.2, if the bound for the weights difference (3.2) holds for all $t \leq t'$ and error bound (3.1) holds for all $t \leq t' - 1$, then (3.1) holds for $t = t'$.*

Proof. We start with analyzing the error $e(t)$ according to

$$\begin{aligned}
e_i(t+1) &= \frac{1}{\sqrt{p}}\beta(t+1)^\intercal\psi(W(t+1)x_i) - y_i \\
&= \frac{1}{\sqrt{p}}\beta(t+1)^\intercal(\psi(W(t+1)x_i) - \psi(W(t)x_i)) + \frac{1}{\sqrt{p}}(\beta(t+1) - \beta(t))^\intercal\psi(W(t)x_i) \\
&\quad + \frac{1}{\sqrt{p}}\beta(t)^\intercal\psi(W(t)x_i) - y_i \\
&= e_i(t) - \frac{\eta}{p}\beta(t+1)^\intercal D_i(t) \sum_{j=1}^{n} D_j(t)bx_j^\intercal x_i e_j(t) - \frac{\eta}{p}\sum_{j=1}^{n} \psi(W(t)x_j)^\intercal\psi(W(t)x_i)e_j(t) \\
&\quad + v_i(t) \\
&= e_i(t) - \eta\sum_{j=1}^{n} \big(H_{ij}(t) + G_{ij}(t)\big)e_j(t) + v_i(t)
\end{aligned}$$

where

$$G_{ij}(t) = \frac{1}{p}\psi(W(t)x_j)^\intercal \psi(W(t)x_i)$$
$$H_{ij}(t) = \frac{x_i^\intercal x_j}{p}\beta(t+1)^\intercal D_i(t)D_j(t)b$$

and $v_i(t)$ is the residual term from the Taylor expansion

$$v_i(t) = \frac{1}{2\sqrt{p}}\sum_{r=1}^{p}\beta_r(t+1)|(w_r(t+1)-w_r(t))^\intercal x_i|^2\psi''(\xi_{ri}(t))$$

with $\xi_{ri}(t)$ between $w_r(t)^\intercal x_i$ and $w_r(t+1)^\intercal x_i$. We can also rewrite the above iteration in vector form as

$$e(t+1) = e(t) - \eta(G(t)+H(t))e(t) + v(t). \tag{A.10}$$

Now for $t = t' - 1$, we wish to show that both $G(t)$ and $H(t)$ are close to their initialization. Notice that

$$\begin{aligned}
|G_{ij}(t) - G_{ij}(0)| &= \frac{1}{p}\Big|\psi(W(t)x_j)^\intercal\psi(W(t)x_i) - \psi(W(t)x_j)^\intercal\psi(W(t)x_i)\Big| \\
&\le \frac{1}{p}\sum_{r=1}^{p}|\psi(w_r(t)^\intercal x_j)||\psi(w_r(t)^\intercal x_i) - \psi(w_r(0)^\intercal x_i)| \\
&\quad + \frac{1}{p}\sum_{r=1}^{p}|\psi(w_r(0)^\intercal x_i)||\psi(w_r(t)^\intercal x_j) - \psi(w_r(0)^\intercal x_j)| \\
&\le c\frac{1}{p}\sum_{r=1}^{p}|w_r(t)^\intercal x_i - w_r(0)^\intercal x_i| + \frac{1}{p}\sum_{r=1}^{p}|w_r(t)^\intercal x_j - w_r(0)^\intercal x_j| \\
&\le c_0\frac{n\sqrt{\log p}}{\gamma\sqrt{p}}(\|x_i\| + \|x_j\|)
\end{aligned}$$

where the second inequality is due to the boundedness of ψ and ψ', and the last inequality is by (3.2). Then we have

$$\|G(t) - G(0)\| \le \max_{j\in[n]}\sum_{i=1}^{n}|G_{ij}(t) - G_{ij}(0)| \le c_0\frac{n^2\sqrt{\log p}}{\gamma\sqrt{p}}. \tag{A.11}$$

For matrix $H(t)$, we similarly have

$$\begin{aligned}
|H_{ij}(t) - H_{ij}(0)| &\le \frac{|x_i^\intercal x_j|}{p}\Big|\beta(t+1)^\intercal D_i(t)D_j(t)b - \beta(0)^\intercal D_i(0)D_j(0)b\Big| \\
&\le \frac{\|x_i\|\|x_j\|}{p}\sum_{r=1}^{p}\Big|b_r\beta_r(t+1)\psi'(w_r(t)^\intercal x_i)\psi'(w_r(t)^\intercal x_j) \\
&\quad - b_r\beta_r(0)\psi'(w_r(0)^\intercal x_i)\psi'(w_r(0)^\intercal x_j)\Big| \\
&\le \frac{|\|x_i\|\|x_j\||}{p}\sum_{r=1}^{p}\Big(|b_r||\beta_r(t+1)-\beta_r(0)||\psi'(w_r(t)^\intercal x_i)\psi'(w_r(t)^\intercal x_j)| \\
&\quad + |b_r||\beta_r(0)||\psi'(w_r(t)^\intercal x_i) - \psi'(w_r(0)^\intercal x_i)||\psi'(w_r(t)^\intercal x_j)| \\
&\quad + |b_r||\beta_r(0)||\psi'(w_r(0)^\intercal x_i)||\psi'(w_r(t)^\intercal x_j) - \psi'(w_r(0)^\intercal x_j)|\Big) \\
&\le c\frac{\|x_i\|\|x_j\|}{p}\sum_{r=1}^{p}\Big(|b_r|\frac{n}{\gamma\sqrt{p}} + |b_r||\beta_r(0)|\frac{n\sqrt{\log p}}{\gamma\sqrt{p}}(\|x_i\| + \|x_j\|)\Big) \\
&\le c_1\frac{n}{\gamma\sqrt{p}} + c_2\frac{n\sqrt{\log p}}{\gamma\sqrt{p}}.
\end{aligned}$$

It follows that

$$\|H(t) - H(0)\| \le \max_{j\in[n]} \sum_{i=1}^{n} |H_{ij}(t) - H_{ij}(0)| \le c_1 \frac{n^2}{\gamma\sqrt{p}} + c_2 \frac{n^2\sqrt{\log p}}{\gamma\sqrt{p}}. \tag{A.12}$$

Next, we bound the residual term $v_i(t)$. Since ψ'' is bounded, we have

$$\begin{aligned}
|v_i(t)| &\le c\frac{1}{\sqrt{p}} \sum_{r=1}^{p} |\beta_r(t+1)| \|w_r(t+1) - w_r(t)\|^2 \\
&\le c\frac{1}{\sqrt{p}} \frac{\eta^2}{p} \sum_{r=1}^{p} |\beta_r(t+1)| \Big(\sum_{i=1}^{n} \|\psi'(w_r(t)^\intercal x_i) b_r x_i e_i(t)\| \Big)^2 \\
&\le c\frac{1}{\sqrt{p}} \frac{\eta^2}{p} \sum_{r=1}^{p} |\beta_r(t+1)| |b_r|^2 \Big(\sum_{i=1}^{n} |e_i(t)| \Big)^2 \\
&\le c\frac{\eta^2 n}{\sqrt{p}} \|e(t)\|^2 \\
&\le c_3 \frac{\eta^2 n\sqrt{n}}{\sqrt{p}} \|e(t)\|.
\end{aligned}$$

This leads to the bound

$$\|v(t)\| = \Big(\sum_{i=1}^{n} |v_i(t)|^2 \Big)^{1/2} \le c_3 \frac{\eta^2 n^2}{\sqrt{p}} \|e(t)\|. \tag{A.13}$$

Combining Eqs. (A.10) to (A.13), we have

$$\begin{aligned}
\|e(t+1)\| &\le \|I_n - \eta(G(t) + H(t))\| \|e(t)\| + \|v(t)\| \\
&\le \Big(\|I_n - \eta G(0)\| + \eta \|G(t) - G(0)\| + \eta \|H(0)\| \\
&\quad + \eta \|H(t) - H(0)\| \Big) \|e(t)\| + \|v(t)\| \\
&\le \Big(1 - \frac{3\eta\gamma}{4} + c_0 \frac{\eta n^2\sqrt{\log p}}{\gamma\sqrt{p}} + \frac{\eta\gamma}{4} + c_1 \frac{\eta n^2}{\gamma\sqrt{p}} + c_2 \frac{\eta n^2\sqrt{\log p}}{\gamma\sqrt{p}} + c_3 \frac{\eta^2 n\sqrt{n}}{\sqrt{p}} \Big) \|e(t)\| \\
&\le (1 - \frac{\eta\gamma}{4}) \|e(t)\|
\end{aligned}$$

where we use Lemma A.3 and $p = \Omega(\frac{n^4 \log p}{\gamma^4})$. □

Proof of Theorem 3.2. We prove the inequality (3.1) by induction. Suppose (3.1) and (3.2) hold for all $t = 1, 2, ..., t' - 1$, by Lemma A.4 and Lemma A.5 we know (3.1) and (3.2) hold for $t = t'$, which completes the proof. □

A.3 Proof of Theorem 4.2

Lemma A.6. *Assume all the inequalities from Lemma A.2 hold. Under the conditions of Theorem 4.2, if the error bound (4.2) holds for all $t = 1, 2, ..., t' - 1$, then*

$$\begin{aligned}
\|w_r(t) - w_r(0)\| &\le c_1 \frac{n\sqrt{\log p}}{\gamma\sqrt{p}} (1 + \eta \tilde{S}_\lambda), \\
|\beta_r(t) - \beta_r(0)| &\le c_2 \frac{n}{\gamma\sqrt{p}} (1 + \eta \tilde{S}_\lambda)
\end{aligned} \tag{A.14}$$

hold for all $t \le t'$, where c_1, c_2 are constants.

Proof. For any $k \le t' - 1$, we apply (4.2) repeatedly on the right hand side of itself to get

$$\|e(k)\| \le \prod_{i=0}^{k-1} \Big(1 - \frac{\eta\gamma}{4} - \eta\lambda(i) \Big) \|e(0)\| + \sum_{i=0}^{k-1} \eta\lambda(i) \prod_{i<j<k} \Big(1 - \frac{\eta\gamma}{4} - \eta\lambda(j) \Big) \|y\|.$$

For $t \leq t' - 1$, we take the sum over $k = 0, .., t$ on both sides of above inequality to obtain

$$\begin{aligned}
\sum_{k=0}^{t} \|e(k)\| &\leq \sum_{k=0}^{t} \prod_{i=0}^{k-1} \Big(1 - \frac{\eta\gamma}{4} - \eta\lambda(i)\Big)\|e(0)\| + \sum_{k=0}^{t}\sum_{i=0}^{k-1} \eta\lambda(i) \prod_{i<j<k} \Big(1 - \frac{\eta\gamma}{4} - \eta\lambda(j)\Big)\|y\| \\
&\leq \sum_{k=0}^{t} \Big(1 - \frac{\eta\gamma}{4}\Big)^{k-1}\|e(0)\| + \sum_{k=0}^{t}\sum_{i=0}^{k-1} \eta\lambda(i)\Big(1 - \frac{\eta\gamma}{4}\Big)^{k-i-1}\|y\| \\
&\leq \sum_{k=0}^{t} \Big(1 - \frac{\eta\gamma}{4}\Big)^{k-1}\|e(0)\| + \eta\|y\| \sum_{k=0}^{t-1} \lambda(i) \sum_{k=i+1}^{T} \Big(1 - \frac{\eta\gamma}{4}\Big)^{k-i-1} \\
&\leq \frac{4}{\eta\gamma}\|e(0)\| + \frac{4}{\gamma}\tilde{S}_\lambda\|y\| \\
&\leq \frac{c\sqrt{n}}{\gamma}(\frac{1}{\eta} + \tilde{S}_\lambda)
\end{aligned}$$

where we use $\|e(0)\| = O(\sqrt{n})$ and $\|y\| = O(\sqrt{n})$. Then for all $t \leq t'$, we have

$$\begin{aligned}
|\beta_r(t) - \beta_r(0)| &\leq \frac{\eta}{\sqrt{p}} \sum_{s=0}^{t-1}\sum_{i=1}^{n} |\psi(w_r(t)x_i)e_i(t)| \\
&\leq c\frac{\eta}{\sqrt{p}} \sum_{s=0}^{t-1}\sum_{i=1}^{n} |e_i(t)| \\
&\leq c\frac{\eta\sqrt{n}}{\sqrt{p}} \sum_{s=0}^{t-1} \|e(t)\| \\
&\leq c\frac{\eta\sqrt{n}}{\sqrt{p}} \frac{\sqrt{n}}{\gamma}(\frac{1}{\eta} + \tilde{S}_\lambda) \\
&\leq c\frac{n}{\gamma\sqrt{p}}(1 + \eta\tilde{S}_\lambda)
\end{aligned}$$

where we use ψ is bounded and (A.6). We also have

$$\begin{aligned}
\|w_r(t) - w_r(0)\| &\leq \frac{\eta}{\sqrt{p}} \sum_{s=0}^{t-1}\sum_{i=1}^{n} \|\psi'(w_r(t)^\intercal x_i)b_r x_i e_i(t)\| \\
&\leq c\frac{\eta}{\sqrt{p}} \sum_{s=0}^{t-1}\sum_{i=1}^{n} |b_r||e_i(t)| \\
&\leq c|b_r|\frac{\eta\sqrt{n}}{\sqrt{p}} \sum_{s=0}^{t-1} \|e(t)\| \\
&\leq c|b_r|\frac{\eta\sqrt{n}}{\sqrt{p}} \frac{\sqrt{n}}{\gamma}(\frac{1}{\eta} + \tilde{S}_\lambda) \\
&\leq c\frac{n\sqrt{\log p}}{\gamma\sqrt{p}}(1 + \eta\tilde{S}_\lambda)
\end{aligned}$$

where we use the fact that ψ' is bounded, (A.6) and (A.7). □

Lemma A.7. *Assume all the inequalities from Lemma A.2 hold. Under the conditions of Theorem 4.2, if the bound for weights difference (A.14) holds for all $t \leq t'$ and error bound (4.2) holds for all $t \leq t' - 1$, then (4.2) holds for $t = t'$.*

Proof. We start by analyzing the error $e(t)$ according to

$$\begin{aligned}
e_i(t+1) &= \frac{1}{\sqrt{p}}\beta(t+1)^\intercal\psi(W(t+1)x_i) - y_i \\
&= \frac{1}{\sqrt{p}}\beta(t+1)^\intercal(\psi(W(t+1)x_i) - \psi(W(t)x_i)) + \frac{1}{\sqrt{p}}(\beta(t+1) - (1-\eta\lambda(t))\beta(t))^\intercal\psi(W(t)x_i) \\
&\quad + (1-\eta\lambda(t))\Big(\frac{1}{\sqrt{p}}\beta(t)^\intercal\psi(W(t)x_i) - y_i\Big) - \eta\lambda(t)y \\
&= (1-\eta\lambda(t))e_i(t) - \frac{\eta}{p}\beta(t+1)^\intercal D_i(t)\sum_{j=1}^n D_j(t)bx_j^\intercal x_i e_j(t) - \frac{\eta}{p}\sum_{j=1}^n \psi(W(t)x_j)^\intercal\psi(W(t)x_i)e_j(t) - \eta\lambda(t)y \\
&\quad + v_i(t) \\
&= (1-\eta\lambda(t))e_i(t) - \eta\sum_{j=1}^n \big(H_{ij}(t) + G_{ij}(t)\big)e_j(t) + v_i(t) - \eta\lambda(t)y
\end{aligned}$$

where

$$\begin{aligned}
G_{ij}(t) &= \frac{1}{p}\psi(W(t)x_j)^\intercal\psi(W(t)x_i) \\
H_{ij}(t) &= \frac{x_i^\intercal x_j}{p}\beta(t+1)^\intercal D_i(t)D_j(t)b
\end{aligned}$$

and $v_i(t)$ is the residual term from a Taylor expansion

$$v_i(t) = \frac{1}{2\sqrt{p}}\sum_{r=1}^p \beta_r(t+1)|(w_r(t+1) - w_r(t))^\intercal x_i|^2\psi''(\xi_{ri}(t))$$

with $\xi_{ri}(t)$ between $w_r(t)^\intercal x_i$ and $w_r(t+1)^\intercal x_i$. We can also rewrite the above iteration in vector form as

$$e(t+1) = (1-\lambda(t))e(t) - \eta(G(t)+H(t))e(t) + v(t) - \eta\lambda(t)y. \tag{A.15}$$

Now for $t = t' - 1$, we show that both $G(t)$ and $H(t)$ are close to their initialization. Using the argument in Lemma A.5, we can obtain following bounds

$$\|G(t) - G(0)\| \le c_1\frac{n^2\sqrt{\log p}}{\gamma\sqrt{p}}(1+\eta\tilde{S}_\lambda) \tag{A.16}$$

$$\|H(t) - H(0)\| \le c_2\frac{n^2\sqrt{\log p}}{\gamma\sqrt{p}}(1+\eta\tilde{S}_\lambda) \tag{A.17}$$

$$\|v(t)\| \le c_3\frac{\eta^2 n^2}{\sqrt{p}}\|e(t)\|. \tag{A.18}$$

Combining Eqs. (A.15) to (A.18), we have

$$\begin{aligned}
\|e(t+1)\| &\le \|(1-\eta\lambda(t))I_n - \eta(G(t)+H(t))\|\|e(t)\| + \|v(t)\| \\
&\le \Big(\|(1-\eta\lambda(t))I_n - \eta G(0)\| + \eta\|G(t) - G(0)\| + \eta\|H(0)\| \\
&\quad + \eta\|H(t) - H(0)\|\Big)\|e(t)\| + \|v(t)\| \\
&\le \Big(1 - \eta\lambda(t) - \frac{3\eta\gamma}{4} + (c_1+c_2)\frac{\eta n^2\sqrt{\log p}}{\gamma\sqrt{p}}(1+\eta\tilde{S}_\lambda) + c_3\frac{\eta^2 n\sqrt{n}}{\sqrt{p}}\Big)\|e(t)\| \\
&\le (1 - \eta\lambda(t) - \frac{\eta\gamma}{4})\|e(t)\|
\end{aligned}$$

where we use Lemma A.3, $p = \Omega(\frac{n^4\log p}{\gamma^4})$ and $\tilde{S}_\lambda = O(\frac{\gamma^2\sqrt{p}}{\eta n^2\sqrt{\log p}})$. □

Proof of Theorem 4.2. We prove the inequality (4.2) by induction. Suppose (4.2) holds for all $t = 1, 2, ..., t'-1$. Then by Lemma A.6 and Lemma A.7 we know (4.2) holds for $t = t'$, which completes the proof. □

B Alignment on Two-Layer Linear Networks

Now we assume $\psi(u) = u$, so that f is a linear network. The loss function with regularization at time t is

$$\mathcal{L}(t, W, \beta) = \frac{1}{2}\Big\|\frac{1}{\sqrt{p}} XW^\intercal\beta - y\Big\|^2 + \frac{1}{2}\lambda(t)\|\beta\|^2. \tag{B.1}$$

The regularized feedback alignment algorithm gives

$$\begin{aligned} W(t+1) &= W(t) - \eta\frac{1}{\sqrt{p}} be(t)^\intercal X \\ \beta(t+1) &= (1-\eta\lambda(t))\beta(t) - \frac{\eta}{\sqrt{p}} W(t)X^\intercal e(t) \end{aligned} \tag{B.2}$$

where $e(t) = \frac{1}{\sqrt{p}} XW(t)^\intercal\beta(t) - y$ is the error vector at time t.

Lemma B.1. *Suppose the network is trained with the regularized feedback alignment algorithm* (B.2)*. Then the prediction error $e(t)$ satisfies the recurrence*

$$e(t+1) = \Big[(1-\eta\lambda(t))I_d - \frac{\eta}{p} XW(0)^\intercal W(0)X^\intercal - \eta\Big(J_1(t) + J_2(t) + J_3(t)\Big)\Big]e(t) - \eta\lambda(t)y \tag{B.3}$$

where

$$\begin{aligned} J_1(t) &= \frac{1}{p} b^\intercal\beta(0)\prod_{i=0}^{t}(1-\eta\lambda(i))XX^\intercal \\ J_2(t) &= -\frac{\eta}{p}\Big(\bar{v}^\intercal X^\intercal\hat{s}(t)XX^\intercal + XX^\intercal s(t-1)\bar{v}^\intercal X^\intercal + X\bar{v}s(t-1)^\intercal XX^\intercal\Big) \\ J_3(t) &= \frac{\eta^2}{p^2}\|b\|^2\Big(\hat{S}(t)XX^\intercal + XX^\intercal s(t-1)s(t-1)^\intercal XX^\intercal\Big) \end{aligned}$$

and

$$\begin{aligned} \bar{v} &= \frac{1}{\sqrt{p}} W(0)^\intercal b \\ s(t) &= \sum_{i=0}^{t} e(i) \\ \hat{s}(t) &= \sum_{i=0}^{t}\prod_{i<k\le t}(1-\eta\lambda(k))e(i) \\ \hat{S}(t) &= \sum_{i=0}^{t}\prod_{i<k\le t}(1-\eta\lambda(k))e(i)^\intercal XX^\intercal\sum_{j=0}^{i-1} e(j). \end{aligned}$$

Proof. We first write $W(t)$ in terms of $W(0)$ and $e(i)$, $i \in [t]$, so that

$$W(t) = W(0) - \frac{\eta}{\sqrt{p}} b\sum_{i=0}^{t-1} e(i)^\intercal X = W(0) - \frac{\eta}{\sqrt{p}} bs(t-1)^\intercal X. \tag{B.4}$$

Similarly, for $\beta(t)$ we have

$$\begin{aligned}
\beta(t) &= \prod_{i=0}^{t-1}(1-\eta\lambda(i))\beta(0) - \frac{\eta}{\sqrt{p}}\sum_{i=0}^{t-1}\prod_{i<k<t}(1-\eta\lambda(k))W(i)X^\intercal e(i) \\
&= \prod_{i=0}^{t-1}(1-\eta\lambda(i))\beta(0) - \frac{\eta}{\sqrt{p}}\sum_{i=0}^{t-1}\prod_{i<k<t}(1-\eta\lambda(k))\Big(W(0) - \frac{\eta}{\sqrt{p}}b\sum_{j=0}^{i-1}e(j)^\intercal X\Big)X^\intercal e(i) \\
&= \prod_{i=0}^{t-1}(1-\eta\lambda(i))\beta(0) - \frac{\eta}{\sqrt{p}}\sum_{i=0}^{t-1}\prod_{i<k<t}(1-\eta\lambda(k))W(0)X^\intercal e(i) \\
&\quad + \frac{\eta^2}{p}b\sum_{i=0}^{t-1}\prod_{i<k<t}(1-\eta\lambda(k))e(i)^\intercal XX^\intercal\sum_{j=0}^{i-1}e(j) \\
&= \prod_{i=0}^{t-1}(1-\eta\lambda(i))\beta(0) - \frac{\eta}{\sqrt{p}}W(0)X^\intercal\hat{s}(t-1) + \frac{\eta^2}{p}b\hat{S}(t-1).
\end{aligned} \tag{B.5}$$

We now study how the error $e(t)$ changes after a single update step, writing

$$\begin{aligned}
e(t+1) &= \frac{1}{\sqrt{p}}XW(t+1)^\intercal\beta(t+1) - y \\
&= \frac{1}{\sqrt{p}}X(W(t+1) - W(t)^\intercal\beta(t+1) + \frac{1}{\sqrt{p}}XW(t)^\intercal(\beta(t+1) - (1-\eta\lambda(t))\beta(t)) \\
&\quad + (1-\eta\lambda(t))\Big(\frac{1}{\sqrt{p}}XW(t)^\intercal\beta(t) - y\Big) - \eta\lambda(t)y \\
&= (1-\eta\lambda(t))e(t) - \frac{\eta}{p}b^\intercal\beta(t+1)XX^\intercal e(t) - \frac{\eta}{p}XW(t)^\intercal W(t)X^\intercal e(t) - \eta\lambda(t)y
\end{aligned}$$

By plugging (B.4) and (B.5) into above equation, we have

$$\begin{aligned}
e(t+1) = (1-\eta\lambda(t))e(t) &\\
&- \frac{\eta}{p}b^\intercal\Big[\prod_{i=0}^{t}(1-\eta\lambda(i))\beta(0) - \frac{\eta}{\sqrt{p}}W(0)X^\intercal\hat{s}(t) + \frac{\eta^2}{p}b\hat{S}(t)\Big]XX^\intercal e(t) \\
&- \frac{\eta}{p}X\Big[W(0) - \frac{\eta}{\sqrt{p}}bs(t-1)^\intercal X\Big]^\intercal\Big[W(0) - \frac{\eta}{\sqrt{p}}bs(t-1)^\intercal X\Big]X^\intercal e(t) \\
&- \eta\lambda(t)y
\end{aligned}$$

After expanding the brackets and rearranging the items, we can obtain (B.3). □

Lemma B.2. *Given $\delta \in (0,1)$ and $\epsilon > 0$, if $p = \Omega(\frac{1}{\epsilon}\log\frac{d}{\delta} + \frac{d}{\epsilon}\log\frac{1}{\epsilon})$, the following inequalities hold with probability at least $1-\delta$*

$$\frac{|b^\intercal\beta(0)|}{\sqrt{p}} \le c\sqrt{\log\frac{1}{\delta}} \tag{B.6}$$

$$\frac{\|b^\intercal W(0)\|}{\sqrt{p}} \le c\sqrt{d\log\frac{d}{\delta}} \tag{B.7}$$

$$\Big|\frac{\|b\|^2}{p} - 1\Big| \le \frac{c}{\sqrt{p}}\sqrt{\log\frac{1}{\delta}} \tag{B.8}$$

$$\Big\|\frac{1}{p}W(0)^\intercal W(0) - I_d\Big\| \le \epsilon \tag{B.9}$$

where c is a constant.

Proof. (B.6) is derived from Lemma C.4. (B.7) is by (B.6) and a union bound argument. (B.8) is by Lemma C.3. (B.9) is by Corollary C.2 □

Proof of Theorem 4.3. We show (4.3) by induction. Assume (4.3) holds for all $t = 0, 1, ..., t'$, we will show it hold for $t = t' + 1$. For any $k \le t'$, we apply (4.3) repeatedly on the right hand side of itself to get

$$\|e(k)\| \le \prod_{i=0}^{k-1}\Big(1 - \frac{\eta\gamma}{2} - \eta\lambda(i)\Big)\|e(0)\| + \sum_{i=0}^{k-1}\eta\lambda(i)\prod_{i<j<k}\Big(1 - \frac{\eta\gamma}{2} - \eta\lambda(j)\Big)\|y\|$$

For $t \le t'$, we take the sum over $k = 0, .., t$ on both sides of above inequality

$$\begin{aligned}
\sum_{k=0}^{t}\|e(k)\| &\le \sum_{k=0}^{t}\prod_{i=0}^{k-1}\Big(1 - \frac{\eta\gamma}{2} - \eta\lambda(i)\Big)\|e(0)\| + \sum_{k=0}^{t}\sum_{i=0}^{k-1}\eta\lambda(i)\prod_{i<j<k}\Big(1 - \frac{\eta\gamma}{2} - \eta\lambda(j)\Big)\|y\| \\
&\le \sum_{k=0}^{t}\Big(1 - \frac{\eta\gamma}{2}\Big)^{k-1}\|e(0)\| + \sum_{k=0}^{t}\sum_{i=0}^{k-1}\eta\lambda(i)\Big(1 - \frac{\eta\gamma}{2}\Big)^{k-i-1}\|y\| \\
&\le \sum_{k=0}^{t}\Big(1 - \frac{\eta\gamma}{2}\Big)^{k-1}\|e(0)\| + \eta\|y\|\sum_{k=0}^{t-1}\lambda(i)\sum_{k=i+1}^{T}\Big(1 - \frac{\eta\gamma}{2}\Big)^{k-i-1} \\
&\le \frac{2}{\eta\gamma}\|e(0)\| + \frac{2}{\gamma}S_\lambda\|y\| \\
&\le \frac{c\sqrt{n}}{\gamma}(\frac{1}{\eta} + S_\lambda)
\end{aligned}$$

where we use $\|e(0)\| = O(\sqrt{n})$ and $\|y\| = O(\sqrt{n})$. With this bound and the inequalities from Lemma B.2, we can bound the norms of $J_1(t)$, $J_2(t)$ and $J_3(t)$ from Lemma B.1. It follows that

$$\|J_1(t)\| \le \frac{1}{p}|b^\intercal\beta(0)|\|XX^\intercal\| \le c\frac{M\sqrt{\log\delta^{-1}}}{\sqrt{p}} \le \frac{\gamma}{16}, \tag{B.10}$$

$$\|J_2(t)\| \le \frac{\eta}{p}\|X\|\|XX^\intercal\|\|\bar{v}\|(2\|s(t-1)\| + \|\hat{s}(t)\|) \le c\frac{\eta}{p}M^{3/2}\sqrt{d\log\frac{d}{\delta}}\frac{\sqrt{n}}{\gamma}(\frac{1}{\eta} + S_\lambda) \le \frac{\gamma}{16} \tag{B.11}$$

and

$$\|J_3(t)\| \le \frac{\eta^2}{p^2}\|b\|^2(\|XX^\intercal\|\|\hat{S}(t)| + \|XX^\intercal\|^2\|s(t-1)\|^2) \le c\frac{\eta^2}{p}M^2\frac{n}{\gamma^2}(\frac{1}{\eta} + S_\lambda)^2 \le \frac{\gamma}{16} \tag{B.12}$$

hold for all $t \le t'$ if $p = \Omega(\frac{Md\log(d/\delta)}{\gamma})$ and $S_\lambda = O(\frac{\gamma\sqrt{\gamma p}}{\eta\sqrt{n}M})$. Furthermore, since $\|\frac{1}{p}W(0)W(0)^\intercal - I_d\| \le \epsilon_0$ with high probability when $p = \Omega(d)$, we have

$$\begin{aligned}
\|\frac{1}{p}XW(0)^\intercal W(0)X^\intercal - \gamma I_d\| &\le \|\frac{1}{p}XW(0)^\intercal W(0)X^\intercal - XX^\intercal\| + \|XX^\intercal - \gamma I_d\| \\
&\le (1+\epsilon)\epsilon_0\gamma + \epsilon\gamma \le \frac{\gamma}{16}
\end{aligned} \tag{B.13}$$

Therefore, combining (B.10), (B.11), (B.12) and (B.3), we have

$$\begin{aligned}
\|e(t'+1)\| &\le \Big(1 - \eta\lambda(t') - \eta\gamma\Big)\|e(t')\| + \eta\Big\|\frac{\eta}{p}XW(0)^\intercal W(0)X^\intercal - \gamma I_d\Big\|\|e(t')\| \\
&\quad + \eta(\|J_1(t')\| + \|J_2(t')\| + \|J_3(t')\|)\|e(t')\| + \eta\lambda(t')\|y\| \\
&\le \Big(1 - \eta\lambda(t') - \eta\gamma\Big)\|e(t')\| + \frac{1}{16}\eta\gamma\|e(t')\| + \frac{3}{16}\eta\gamma\|e(t')\| + \eta\lambda(t')\|y\| \\
&\le \Big(1 - \eta\lambda(t') - \frac{\eta\gamma}{2}\Big)\|e(t')\| + \eta\lambda(t')\|y\|
\end{aligned}$$

which completes the proof. □

Proof of Proposition 4.5. By Corollary C.2, if $d = \Omega(\frac{1}{\epsilon}\log\frac{n}{\delta} + \frac{n}{\epsilon}\log\frac{1}{\epsilon})$, we have

$$\|XX^\intercal - I_n\| \le \epsilon$$

It follows that $\lambda_{\min}(XX^\intercal) \ge 1 - \epsilon$ and $\lambda_{\max}(XX^\intercal) \le 1 + \epsilon \le (1 + 4\epsilon)(1 - \epsilon)$ for $\epsilon < 1/2$. □

Lemma B.3. *Recall from Lemma B.1 that*

$$\beta(t) = \prod_{i=0}^{t-1}(1-\eta\lambda(i))\beta(0) - \frac{\eta}{\sqrt{p}}W(0)X^\intercal\hat{s}(t-1) + \frac{\eta^2}{p}b\hat{S}(t-1)$$

with $\hat{s}(t) = \sum_{i=0}^{t}\prod_{i<k\le t}(1-\eta\lambda(k))e(i)$ *and* $\hat{S}(t) = \sum_{i=0}^{t}\prod_{i<k\le t}(1-\eta\lambda(k))e(i)^\intercal XX^\intercal\sum_{j=0}^{i-1}e(j)$. *Under the conditions of Theorem 4.6, if* $t > C_1\frac{\log(p/\eta)}{\eta\lambda}$ *and* $\hat{S}(t) \ge \max(C_2\frac{\sqrt{p\gamma}}{\eta}\|\hat{s}(t)\|, 1)$ *for some positive constants* C_1 *and* C_2, *then* $\cos\angle(b,\beta(t)) \ge c$ *for some constant* $c = c_\delta$.

Proof. We compute the cosine of the angle between $\beta(t)$ and b. With probability $1-\delta$,

$$\begin{aligned}
\cos\angle(b,\beta(t)) &= \frac{b^\intercal\beta(t)}{\|b\|\|\beta(t)\|} = \frac{\frac{b}{\|b\|}^\intercal\beta(t)}{\|\beta(t)\|} \\
&\ge \frac{\frac{\eta^2}{p}\|b\|\hat{S}(t-1) - (1-\eta\lambda)^t\|\beta(0)\| - \frac{\eta}{\sqrt{p}}\|\frac{b}{\|b\|}^\intercal W(0)\|\|X\|\|\hat{s}(t-1)\|}{\frac{\eta^2}{p}\|b\|\hat{S}(t-1) + (1-\eta\lambda)^t\|\beta(0)\| + \frac{\eta}{\sqrt{p}}\|W(0)\|\|X\|\|\hat{s}(t-1)\|} \\
&\ge \frac{c_1'\frac{\eta^2}{\sqrt{p}}\hat{S}(t-1) - c_2'\sqrt{p}(1-\eta\lambda)^t - c_3'\eta\sqrt{\frac{d\gamma}{p}}\|\hat{s}(t-1)\|}{c_1'\frac{\eta^2}{\sqrt{p}}\hat{S}(t-1) + c_2'\sqrt{p}(1-\eta\lambda)^t + c_4'\eta\sqrt{\gamma}\|\hat{s}(t-1)\|}
\end{aligned}$$

where we use (B.8), (B.9) and the tail bound for standard Gaussian vectors, and c_i' are constants that only depend on δ. Notice that if $t = \Omega(\frac{\log(p/\eta)}{\eta\lambda})$, we have $c_2'\sqrt{p}(1-\eta\lambda)^t = O(\frac{\eta^2}{\sqrt{p}})$. It follows that $\cos\angle(b,\beta(t)) \ge c$ if $\hat{S}(t-1) = \Omega(\frac{\sqrt{p\gamma}}{\eta}\|\hat{s}(t-1)\| + 1)$. □

Lemma B.4. *Consider the orthogonal decomposition* $e(t) = a(t)\bar{y} + \xi(t)$, *where* $\bar{y} = -y/\|y\|$ *and* $\xi(t) \perp y$. *Under the conditions of Theorem 4.6, there exists a constant* $C_\tau > 0$ *such that for any* $t \in [\tau, T]$ *with* $\tau = \frac{C_\tau}{\eta\lambda}$, *we have*

$$a(t) \ge \frac{\lambda-\gamma}{\lambda+\gamma}\|y\| \tag{B.14}$$

and

$$\|\xi(t)\| \le \frac{\gamma}{\lambda+\gamma}\|y\|. \tag{B.15}$$

Proof. By Theorem 4.3, we have for all $t \le T$, $\|e(t)\| \le (1-\eta\lambda-\eta\gamma/2)\|e(t)\| + \eta\lambda\|y\|$. By rearranging the terms, we have

$$\|e(t+1)\| - \frac{\lambda}{\lambda-\gamma/2}\|y\| \le (1-\eta\lambda-\frac{\eta\gamma}{2})\Big(\|e(t)\| - \frac{\lambda}{\lambda-\gamma/2}\|y\|\Big)$$

or

$$\|e(t)\| - \frac{\lambda}{\lambda-\gamma/2}\|y\| \le (1-\eta\lambda-\frac{\eta\gamma}{2})^t\Big(\|e_0\| - \frac{\lambda}{\lambda-\gamma/2}\|y\|\Big) \le (1-\eta\lambda)^t(\|e_0\| + \|y\|).$$

Notice that $\|y\|$ and $\|e(0)\|$ are of the same order, so when $t \in [\tau_1, T]$ with $\tau_1 = \frac{c_1}{\eta\lambda}$ and some constant c_1, we have

$$\|e(t)\| \le \frac{\lambda+\gamma/2}{\lambda-\gamma/2}\|y\|. \tag{B.16}$$

In order to get a lower bound for $a(t)$, we multiply $\bar{y}^\intercal$ on both sides of (B.3). It follows that for $t \in [\tau_1, T]$

$$\begin{aligned}
a(t+1) &\ge \bar{y}^\intercal\Big(1-\eta\lambda-\eta\gamma\Big)e(t) - \eta\|\frac{1}{p}XW(0)^\intercal W(0)X^\intercal - \gamma I_d\|\|e(t)\| \\
&\quad - \eta(\|J_1(t)\| + \|J_2(t)\| + \|J_3(t)\|)\|e(t)\| + \eta\lambda\|y\| \\
&\ge (1-\eta\lambda-\eta\gamma)a(t) - \frac{1}{4}\eta\gamma\|e(t)\| + \eta\lambda\|y\| \\
&\ge (1-\eta\lambda-\eta\gamma)a(t) + \frac{1}{2}\eta\gamma\|y\|.
\end{aligned}$$

In the second inequality, we use the bounds (B.10), (B.11), (B.12) and (B.13). The last inequality is by (B.16) and $\lambda \geq 3\gamma$. Following a similar derivation, we have

$$a(t) - \frac{\lambda - \gamma/2}{\lambda + \gamma}\|y\| \geq (1 - \eta\lambda - \eta\gamma)^{t-\tau_1}\Big(a(\tau_1) - \frac{\lambda - \gamma/2}{\lambda + \gamma}\|y\|\Big) \geq -(1 - \eta\lambda)^{t-\tau_1}(\|e(\tau_1)\| + \|y\|).$$

The bound (B.14) holds when $t \in [\tau_1 + \tau_2, T]$ with $\tau_2 = \frac{c_2}{\eta\lambda}$ and some constant c_2. Then we multiply $\frac{\xi(t+1)^\intercal}{\|\xi(t+1)\|}$ on both sides of (B.3). This establishes that for $t \in [\tau_1, T]$

$$\begin{aligned}
\|\xi(t+1)\| &\leq \frac{\xi(t+1)^\intercal}{\|\xi(t+1)\|}\Big(1 - \eta\lambda - \eta\gamma\Big)e(t) + \eta\|\frac{1}{p}XW(0)^\intercal W(0)X^\intercal - \gamma I_d\|\|e(t)\| \\
&\quad + \eta(\|J_1(t)\| + \|J_2(t)\| + \|J_3(t)\|)\|e(t)\| + \eta\lambda\|y\| \\
&\leq (1 - \eta\lambda - \eta\gamma)\|\xi(t)\| + \frac{\eta\gamma}{4}\|e(t)\| \\
&\leq (1 - \eta\lambda - \eta\gamma)\|\xi(t)\| + \frac{\eta\gamma}{2}\eta\gamma\|y\|.
\end{aligned}$$

The first inequality is by $\xi(t+1)^\intercal y = 0$ and in the second inequality we use $\xi(t+1)^\intercal e(t) = \xi(t+1)^\intercal \xi(t) \leq \|\xi(t+1)\|\|\xi(t)\|$. It follows that

$$\|\xi(t)\| - \frac{\gamma/2}{\lambda + \gamma}\|y\| \leq (1 - \eta\lambda - \eta\gamma)^{t-\tau_1}\Big(\|\xi(0)\| - \frac{\gamma/2}{\lambda + \gamma}\|y\|\Big) \leq (1 - \eta\lambda)^{t-\tau_1}(\|e(\tau_1)\| + \|y\|).$$

The bound (B.15) holds when $t \in [\tau_1 + \tau_3, T]$ with $\tau_3 = \frac{c_3}{\eta\lambda}$ for a constant c_3. Finally, the bounds (B.14) and (B.15) hold when $t \in [\tau, T]$ with $\tau = \tau_1 + \max(\tau_2, \tau_3)$. □

Lemma B.5. *Under the conditions of Theorem 4.6, suppose* $T = \lfloor \frac{S_\lambda}{\lambda} \rfloor = C_T \frac{\sqrt{p}}{\eta\sqrt{n\gamma}}$. *Then we have* $\hat{S}(T) \geq \tilde{c}\frac{\sqrt{p\gamma}}{\eta}\|\hat{s}(T)\|$, *where* C_T *and* $\tilde{c}$ *are positive constants.*

Proof. Notice that

$$e(i)^\intercal XX^\intercal e(j) \geq \gamma e(i)^\intercal e(j) - \|e(i)\|\|e(j)\|\|XX^\intercal - \gamma I\| \geq \gamma e(i)^\intercal e(j) - \epsilon\gamma\|e(i)\|\|e(j)\|.$$

For $i \in [T/2, T]$ and τ defined in Lemma B.4, we have

$$\begin{aligned}
e(i)^\intercal XX^\intercal \sum_{j<i} e(j) &= e(i)^\intercal XX^\intercal \sum_{\tau \leq j < i} e(j) + e(i)^\intercal XX^\intercal \sum_{j<\tau} e(j) \\
&\geq \sum_{\tau \leq j < i}\Big(\gamma e(i)^\intercal e(j) - \epsilon\gamma\|e(i)\|\|e(j)\|\Big) - 2\gamma\sum_{j<\tau}\|e(i)\|\|e(j)\| \\
&\geq \sum_{\tau \leq j < i}\gamma\Big(a(i)a(j) - \|\xi(i)\|\|\xi(j)\| - \epsilon\|e(i)\|\|e(j)\|\Big) - 2c\tau\gamma\|y\|^2 \\
&\geq (i - \tau)\gamma\Big[\Big(\frac{\lambda - \gamma}{\lambda + \gamma}\Big)^2\|y\|^2 - \Big(\frac{\gamma}{\lambda + \gamma}\Big)^2\|y\|^2 - \epsilon\Big(\frac{\lambda + \gamma/2}{\lambda - \gamma/2}\Big)^2\|y\|^2 - \frac{2c\tau}{i - \tau}\|y\|^2\Big] \\
&\geq \frac{T}{8}\gamma\|y\|^2 = \frac{C_T}{8}\frac{\sqrt{p}}{\eta\sqrt{n\gamma}}\gamma\|y\|^2 \\
&\geq c\frac{\sqrt{p\gamma}}{\eta}\|y\|.
\end{aligned} \tag{B.17}$$

The second inequality is the orthogonal decomposition of $e(i)$ and $\|e(i)\| \leq c\|y\|$ given by (4.3). The third inequality is by (B.14), (B.15) and (B.16) from Lemma B.4. The fourth inequality is by $\lambda = \Omega(\gamma)$, $i - \tau \geq T/4$ and the fact that $\tau/(i - \tau)$ is small ($p = \Omega(n)$). The last inequality is by

$\|y\| = \Theta(\sqrt{n})$. Therefore,

$$\begin{aligned}
\hat{S}(T) &= \sum_{i=0}^{T}(1-\eta\lambda)^{T-i}e(i)^\intercal XX^\intercal \sum_{j<i} e(j) \\
&= \sum_{i=T/2}^{T}(1-\eta\lambda)^{T-i}e(i)^\intercal XX^\intercal \sum_{j<i} e(j) + (1-\eta\lambda)^{T/2}\sum_{i=0}^{T/2}(1-\eta\lambda)^{T/2-i}e(i)^\intercal XX^\intercal \sum_{j<i} e(j) \\
&\geq \sum_{i=T/2}^{T}(1-\eta\lambda)^{T-i}c\frac{\sqrt{p\gamma}}{\eta}\|y\| + (1-\eta\lambda)^{T/2}\sum_{i=0}^{T/2}(1-\eta\lambda)^{T/2-i}c'T\gamma\|y\|^2 \\
&\geq \frac{c}{2}\frac{\sqrt{p\gamma}}{\eta}\frac{\|y\|}{\eta\lambda} - (1-\eta\lambda)^{T/2}\frac{c'T\gamma\|y\|^2}{\eta\lambda} \\
&\geq \frac{c}{4}\frac{\sqrt{p\gamma}}{\eta}\frac{\|y\|}{\eta\lambda}
\end{aligned}$$

where the last inequality is by $(1-\eta\lambda)^{T/2} \ll 1$ when $p = \Omega(n)$. On the other hand,

$$\|\hat{s}(T)\| \leq \sum_{i=0}^{T}(1-\eta\lambda)^{T-i}\|e(i)\| \leq \frac{c}{\eta\lambda}\|y\|.$$

Combining the above inequalities gives the proof. □

Proof of Theorem 4.6. First, notice that $\lambda(t) = 0$ when $t > T$. By Theorem 4.3 we have that the prediction error converges to zero exponentially fast, or $\|e(t+1)\| \leq (1-\eta\gamma/2)\|e(t)\|$. It follows that $\hat{S}(t) \to \hat{S}(\infty)$ and $\hat{s}(t) \to \hat{s}(\infty)$ as $t \to \infty$. By Lemma B.3, we know it suffices to show $\hat{S}(\infty) \geq C\frac{\sqrt{p\gamma}}{\eta}\|\hat{s}(\infty)\|$ with some constant C. Since

$$\hat{S}(\infty) = \sum_{i=0}^{\infty}(1-\eta\lambda)^{(T-i)_+}e(i)^\intercal XX^\intercal \sum_{j<i} e(j) = \hat{S}(T) + \sum_{i>T} e(i)^\intercal XX^\intercal \sum_{j<i} e(j)$$

and

$$\hat{s}(\infty) = \sum_{i=0}^{\infty}(1-\eta\lambda)^{(T-i)_+}e(i) = \hat{s}(T) + \sum_{i>T} e(i),$$

by Lemma B.5, it suffices to show

$$\sum_{i>T} e(i)^\intercal XX^\intercal \sum_{j<i} e(j) \geq C\frac{\sqrt{p\gamma}}{\eta}\sum_{i>T}\|e(i)\|. \tag{B.18}$$

We write $g = XX^\intercal \sum_{j<T} e(j)$. Then we have

$$\begin{aligned}
\|g\| &\geq \lambda_{\min}(XX^\intercal)\Big[\Big\|\sum_{\tau\geq j<T} e(j)\Big\| - \sum_{j<\tau}\|e(j)\|\Big] \\
&\geq \lambda_{\min}(XX^\intercal)\Big[\sum_{\tau\geq j<T} a(j) - \sum_{j<\tau}\|e(j)\|\Big] \\
&\geq \gamma\Big[(T-\tau)\Big(\frac{\lambda-\gamma}{\lambda+\gamma}\Big)\|y\| - \tau c\|y\|\Big]
\end{aligned} \tag{B.19}$$

and

$$\begin{aligned}
\|g\| &\leq \|XX^\intercal\|\Big(\sum_{j<\tau}\|e(j)\| + \sum_{\tau\geq j<T}\|e(j)\|\Big) \\
&\leq (1+\epsilon)\gamma\Big[\tau c\|y\| + (T-\tau)\Big(\frac{\lambda+\gamma/2}{\lambda-\gamma/2}\Big)\|y\|\Big]
\end{aligned} \tag{B.20}$$

where we use the bounds (B.14) and (B.16) from Lemma B.4. We further denote $\alpha(t) = \bar{g}^\intercal e(t)$ where $\bar{g} = g/\|g\|$. Following the same calculation in (B.17), we have

$$\begin{aligned} g^\intercal e(T) &= e(T)^\intercal XX^\intercal \sum_{j<T} e(j) \\ &\geq (T-\tau)\gamma\Big[\Big(\frac{\lambda-\gamma}{\lambda+\gamma}\Big)^2\|y\|^2 - \Big(\frac{\gamma}{\lambda+\gamma}\Big)^2\|y\|^2 - \epsilon\Big(\frac{\lambda+\gamma/2}{\lambda-\gamma/2}\Big)^2\|y\|^2 - \frac{2c\tau}{T-\tau}\|y\|^2\Big]. \end{aligned}$$

Then

$$\begin{aligned} \frac{\alpha(T)}{\|e(T)\|} &\geq \frac{g^\intercal e(T)}{\|g\|\|e(T)\|} \\ &\geq \frac{(T-\tau)\gamma\Big[\Big(\frac{\lambda-\gamma}{\lambda+\gamma}\Big)^2\|y\|^2 - \Big(\frac{\gamma}{\lambda+\gamma}\Big)^2\|y\|^2 - \epsilon\Big(\frac{\lambda+\gamma/2}{\lambda-\gamma/2}\Big)^2\|y\|^2 - \frac{2c\tau}{T-\tau}\|y\|^2\Big]}{(1+\epsilon)\gamma\Big[\tau c\|y\| + (T-\tau)\Big(\frac{\lambda+\gamma/2}{\lambda-\gamma/2}\Big)\|y\|\Big] \times \Big(\frac{\lambda+\gamma/2}{\lambda-\gamma/2}\Big)\|y\|} \\ &\geq \frac{\Big[\Big(\frac{\lambda-\gamma}{\lambda+\gamma}\Big)^2 - \Big(\frac{\gamma}{\lambda+\gamma}\Big)^2 - \epsilon\Big(\frac{\lambda+\gamma/2}{\lambda-\gamma/2}\Big)^2 - \frac{2c\tau}{T-\tau}\Big]}{(1+\epsilon)\Big[\frac{\tau c}{T-\tau} + \Big(\frac{\lambda+\gamma/2}{\lambda-\gamma/2}\Big)\Big] \times \Big(\frac{\lambda+\gamma/2}{\lambda-\gamma/2}\Big)}. \end{aligned}$$

Notice that $T/\tau = \Omega(\sqrt{p/n})$, so that when p/n, λ/γ are large and ϵ is small, we have

$$\alpha(T) \geq \frac{3}{4}\|e(T)\|. \tag{B.21}$$

In order to obtain the lower bound on $\alpha(t)$ for all $t \geq T$, we multiply $\bar{g}^\intercal$ on both sides of (B.3). Notice $\lambda(t) = 0$ and apply the bounds (B.10), (B.11), (B.12) and (B.13). We have that

$$\begin{aligned} \alpha(t+1) &\geq (1-\eta\gamma)\bar{g}^\intercal e(t) - \eta\|\frac{1}{p}XW(0)^\intercal W(0)X^\intercal - \gamma I_d\|\|e(t)\| \\ &\quad - \eta(\|J_1(t)\| + \|J_2(t)\| + \|J_3(t)\|)\|e(t)\| \\ &\geq (1-\eta\gamma)\alpha(t) - \frac{\eta\gamma}{4}\|e(t)\| \end{aligned}$$

or for $t \geq T$,

$$\alpha(t) \geq (1-\eta\gamma)^{t-T}\alpha(T) - \frac{\eta\gamma}{4}\sum_{i=T}^{t-1}(1-\eta\gamma)^{t-i}\|e(i)\|. \tag{B.22}$$

Taking the sum over $t > T$, we have

$$\begin{aligned} \sum_{t>T}\alpha(t) &\geq \sum_{t>T}(1-\eta\gamma)^{t-T}\alpha(T) - \frac{\eta\gamma}{4}\sum_{t>T}\sum_{i=T}^{t-1}(1-\eta\gamma)^{t-i}\|e(i)\| \\ &\geq \frac{1-\eta\gamma}{\eta\gamma}\alpha(T) - \frac{\eta\gamma}{4}\sum_{i>T}\|e(i)\|\sum_{t>i}(1-\eta\gamma)^{t-i} \\ &\geq \frac{1-\eta\gamma}{\eta\gamma}\Big(\alpha(T) - \frac{\eta\gamma}{4}\sum_{i>T}\|e(i)\|\Big) \\ &\geq \frac{1-\eta\gamma}{\eta\gamma}(\alpha(T) - \frac{1}{2}\|e(T)\|) \\ &\geq \frac{1-\eta\gamma}{4\eta\gamma}\|e(T)\|. \end{aligned} \tag{B.23}$$

The second inequality follows from switching the order of sums. The fourth inequality is by exponential convergence after T steps. The last inequality is by (B.21). With the above inequalities, we

are ready to bound the left hand side of (B.18), obtaining

$$\begin{aligned}\sum_{i>T} e(i)^\intercal XX^\intercal \sum_{j<i} e(j) &= \sum_{i>T} e(i)^\intercal XX^\intercal \sum_{j<T} e(j) + \sum_{i>T} e(i)^\intercal XX^\intercal \sum_{j\geq T} e(j) \\ &\geq \sum_{t>T} \alpha(t)\|g\| - 2\gamma\Big(\sum_{i\geq t}\|e(i)\|\Big)^2 \\ &\geq \frac{1-\eta\gamma}{4\eta\gamma}\|e(T)\|\gamma\Big[(T-\tau)\Big(\frac{\lambda-\gamma}{\lambda+\gamma}\Big)\|y\| - \tau c\|y\|\Big] - 2\gamma\frac{4}{\eta^2\gamma^2}\|e(T)\|^2 \\ &\geq \frac{1-\eta\gamma}{4\eta\gamma}\|e(T)\|\gamma\Big[(T-\tau)\Big(\frac{\lambda-\gamma}{\lambda+\gamma}\Big)\|y\| - \tau c\|y\| - \frac{64}{\eta\gamma(1-\eta\gamma)}\|y\|\Big] \\ &\geq \frac{1-\eta\gamma}{4\eta\gamma}\|e(T)\|\gamma\frac{T}{2}\|y\| = \frac{1-\eta\gamma}{4\eta\gamma}\|e(T)\|\gamma\frac{C_T}{2}\frac{\sqrt{p}}{\eta\sqrt{n\gamma}}\|y\| \\ &\geq C\frac{1-\eta\gamma}{4\eta\gamma}\frac{\sqrt{p\gamma}}{\eta}\|e(T)\|.\end{aligned} \tag{B.24}$$

The second inequality is by (B.23) and (B.19). The third inequality is by $\|e(T)\| \leq 2\|y\|$. The last inequality is by $\|y\| = \Theta(\sqrt{n})$. On the other hand,

$$\sum_{i>T}\|e(i)\| \leq \sum_{i>T}(1-\eta\gamma/2)^{i-T}\|e(T)\| = \frac{1-\eta\gamma/2}{\eta\gamma/2}\|e(T)\| \tag{B.25}$$

Combining (B.24) and (B.25) implies (B.18), as desired. □

C Technical Lemmas

In this section, we list technical lemmas that are used in our proofs, with references. The first is a variant of the Restricted Isometry Property that bounds the spectral norm of a random Gaussian matrix around 1 with high probability.

Lemma C.1 (Hand & Voroninski, 2018)**.** *Let $A \in \mathbb{R}^{m\times n}$ has i.i.d. $\mathcal{N}(0, 1/m)$ entries. Fix $0 < \varepsilon < 1$, $k < m$, and a subspace $T \subseteq \mathbb{R}^n$ of dimension k, then there exists universal constants c_1 and γ_1, such that with probability at least $1 - (c_1/\varepsilon)^k e^{-\gamma_1\varepsilon m}$,*

$$(1-\varepsilon)\|v\|_2^2 \leq \|Av\|_2^2 \leq (1+\varepsilon)\|v\|_2^2, \quad \forall v \in T.$$

Let us take $k = n$ in Lemma C.1 to get the following corollary.

Corollary C.2. *Let $A \in \mathbb{R}^{m\times n}$ has i.i.d. $\mathcal{N}(0, 1/m)$ entries. For any $0 < \varepsilon < 1$, there exists universal constants c_2 and γ_2, such that with probability at least $1 - (c_2/\varepsilon)^d e^{-\gamma_2\varepsilon m}$,*

$$\|A^\intercal A - I_m\| \leq \varepsilon$$

Then following lemma gives tail bounds for χ^2 random variables.

Lemma C.3 (Laurent & Massart, 2000)**.** *Suppose $X \sim \chi^2_p$, then for all $t \geq 0$ it holds*

$$\mathbb{P}\{X - p \geq 2\sqrt{pt} + 2t\} \leq e^{-t}$$

and

$$\mathbb{P}\{X - p \leq -2\sqrt{pt}\} \leq e^{-t}.$$

For two independent random Gaussian vectors, their inner product can be controlled with the following tail bound.

Lemma C.4 (Gao & Lafferty, 2020)**.** *Let $X, Y \in \mathbb{R}^p$ be independent random Gaussian vectors where $X_r \sim \mathcal{N}(0, 1)$ and $Y_r \sim \mathcal{N}(0, 1)$ for all $r \in [p]$, then it holds*

$$\mathbb{P}(|X^\intercal Y| \geq \sqrt{2pt} + 2t) \leq 2e^t.$$