Layer-Wise Feedback Alignment is Conserved in Deep Neural Networks

Zachary Robertson [1] **Oluwasanmi Koyejo** [1]

Abstract

In the quest to enhance the efficiency and bio-plausibility of training deep neural networks, Feedback Alignment (FA), which replaces the backward pass weights with random matrices in the training process, has emerged as an alternative to traditional backpropagation. While the appeal of FA lies in its circumvention of computational challenges and its plausible biological alignment, the theoretical understanding of this learning rule remains partial. This paper uncovers a set of conservation laws underpinning the learning dynamics of FA, revealing intriguing parallels between FA and Gradient Descent (GD). Our analysis reveals that FA harbors implicit biases akin to those exhibited by GD, challenging the prevailing narrative that these learning algorithms are fundamentally different. Moreover, we demonstrate that these conservation laws elucidate sufficient conditions for layer-wise alignment with feedback matrices in ReLU networks. We further show that this implies over-parameterized two-layer linear networks trained with FA converge to minimum-norm solutions. The implications of our findings offer avenues for developing more efficient and biologically plausible alternatives to backpropagation through an understanding of the principles governing learning dynamics in deep networks.

1. Introduction

Backpropagation, a widely successful learning rule for artificial neural networks, has been instrumental in advancing deep learning. Nevertheless, it presents significant challenges, including the intricate backward pass mechanism, which complicates training parallelism due to communication bottlenecks. In addition, it lacks biological plausibility, further limiting its practical utility.

[1]Department of Computer Science, Stanford, United States. Correspondence to: Zachary Robertson <zroberts@stanford.edu>.

In *ICML Workshop on Localized Learning (LLW)*, Honolulu, Hawaii, USA. 2023.

As an alternative, Feedback Alignment (FA) offers an attractive learning rule that alleviates the computational complexities and bio-plausibility issues associated with backpropagation (Lillicrap et al., 2016). FA replaces the backward pass with random feedback matrices, promising a more straightforward approach to training neural networks.

Despite its advantages, the understanding of FA's underlying principles, particularly in the context of deep neural networks, remains elusive. This gap in knowledge motivates our current investigation, which seeks to unravel the inherent laws that govern the learning dynamics under FA.

Our work makes several significant contributions. Firstly, we establish a set of conservation laws for the learning dynamics under FA, elucidating the implicit bias exhibited by this learning rule—a bias distinct from that of backpropagation. Secondly, these conservation laws enable us to identify a sufficient condition for layer-wise alignment with feedback matrices. Lastly, we provide evidence that two-layer linear networks trained with FA converge to a global optimum.

We believe that our results will have broad implications for future research and practical applications. By quantifying the properties of alternative learning rules like FA, our analysis provides valuable insights that can inform both theoretical advancements and the design of more efficient and biologically plausible alternatives to backpropagation.

2. Related Work

The pioneering work by Stork (1989) questioned the biological plausibility of backpropagation, leading to the exploration of alternative learning rules. Feedback Alignment (FA) emerged as a promising candidate in this regard. Lillicrap et al. (2016) first introduced FA as an effective learning rule for deep neural networks. This approach was further investigated by Nøkland (2016), who demonstrated that random synaptic feedback weights could support error backpropagation for deep learning.

The dynamics of learning with FA have also been the subject of several studies (Refinetti et al., 2021; Song et al., 2021; Launay et al., 2020; Lechner, 2020; Bordelon & Pehlevan, 2022). For instance, Refinetti et al. (2021) highlighted the dynamics of aligning before memorizing in learning with

FA, while a parallel line of work by Launay et al. (2020) showcased that FA can be successfully scaled to modern deep learning tasks and architectures. The work of Lechner (2020) also shows alignment of the learning rule updates with the gradient for single-output networks trained with a variant of feedback alignment, but there are counter-examples for the multi-output setting. There have also been works investigating bio-plausible learning rules applied to networks with large width (Song et al., 2021; Bordelon & Pehlevan, 2022). Our analysis can be seen as a generalization of the approach taken by Lillicrap et al. (2016) to investigate convergence of FA. In particular, our Theorem 5.1 is inspired by a relation satisfied by the parameter updates used in their main convergence result for linear feedback alignment.

Finally, our study is related to the implicit bias of models trained with gradient descent (Vallet et al., 1989; Duin, 2000; Du et al., 2018; Soudry et al., 2018; Belkin et al., 2019). There is a longer history of investigating the phenomena of gradient descent picking solutions that generalize well in linear models (Vallet et al., 1989; Duin, 2000). However, more recently, there has been work looking into similar phenomena involving deep neural networks (Du et al., 2018; Soudry et al., 2018; Belkin et al., 2019). In particular, a related conservation law of neural networks was studied by Du et al. (2018), where they examined the self-balancing property of layers in deep homogeneous models. While their work provides valuable insights into the role of conservation laws in learning dynamics, it does not directly address the implications on alignment as a result of these laws.

Our work diverges from this existing body of literature by proposing a set of conservation laws specifically tailored to the learning dynamics under FA. A unique feature of our study is that these conservation laws yield immediate implications on alignment as a corollary, offering a more comprehensive understanding of the FA's underlying principles. To the best of our knowledge, this is the first result establishing layer-wise alignment for a non-linear network trained with feedback alignment, paving the way for future research in layer-wise alignment in more general settings such as wide neural networks or for other learning rules.

3. Basic Notation

To fully comprehend the dynamics underpinning Feedback Alignment (FA), we first need to establish the appropriate mathematical notation and formalism. This section will clarify the key definitions and operations used throughout this study. We will denote scalars and vectors by lowercase letters (e.g., x, y, z) and matrices by uppercase letters (e.g., A, B, C). The symbol $\odot$ represents the Hadamard (element-wise) product of two matrices or vectors of the same dimensions. Importantly, we denote the trace of a square matrix A by $Tr(A)$. Finally, we note that the trace operator can be used to compute the inner product between two matrices A and B of the same dimension. We denote this inner product as $\langle A, B\rangle = Tr(A^T B)$, where A^T is the transpose of A.

4. Feedback Alignment: A Closer Look

Consider a neural network f parameterized with L layers, where each layer is denoted as l, with $l \in 1, 2, \ldots, L$. Each layer is associated with an activation function ϕ and a weight matrix $W_l \in \mathbb{R}^{n_{l+1} \times n_l}$. Here, n_l refers to the number of neurons in layer l.

For a given input $x \in \mathbb{R}^{n_0}$, the pre-activation h^l and the activation a^l of layer l are computed recursively as follows:

$$h_l = W_l a_{l-1},\ a_l = \phi(h_l)$$

where ϕ denotes the non-linear activation function, a_{l-1} is the activation of the previous layer, and $a_0 = x$.

In FA, the feedback weights, denoted as $B_l \in \mathbb{R}^{n_l \times n_{l+1}}$, are fixed random matrices independent of W_l. The error δ_l at layer l is calculated as:

$$\delta_l = \phi'(h_l) \odot B_{l+1}\delta_{l+1},\ \delta_L = \nabla_{a_L}\mathcal{L}(f)$$

where $\mathcal{L}(f)$ is the loss function applied to the network, and $\nabla_{a_L}\mathcal{L}$ is the gradient of $\mathcal{L}$ with respect to the final layer activation a_L.

The weight update under the FA rule is then given by:

$$\Delta W_l = -\eta \cdot (a_l)^\top \delta_{l+1}^\top$$

where η is the learning rate.

We note that unlike in backpropagation, the feedback matrices B_l are not tied to the feedforward weights W_l. In that setting, we would have time-dependent feedback matrices $B_l(t) = W_l(t)$. This introduces an alignment challenge since there is no guarantee the feedback matrices align with the learned weights.

5. Theoretical Analysis

Building upon the FA formalism detailed in the previous section, we aim to present our main results and provide insight on their distinctive characteristics and how they differ from the traditional backpropagation (BP) in the context of learning rules.

It has been previously observed that the matrices learned through feedback alignment tend to align with their respective feedback matrices (Lillicrap et al., 2016; Nøkland, 2016;

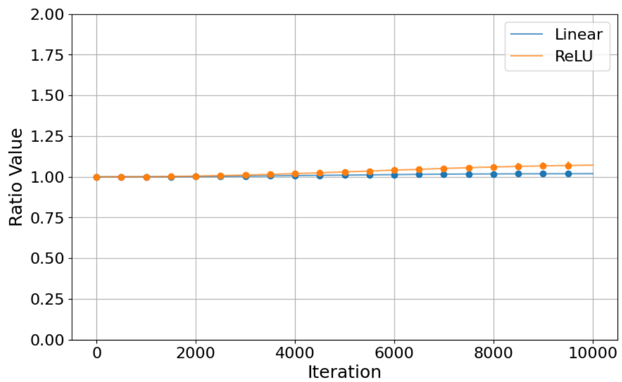

Figure 1. We plot the ratio $\frac{\langle W_2(t), B_2\rangle - \frac{1}{2}\|W_1(t)\|_F^2}{\langle W_2(0), B_2\rangle - \frac{1}{2}\|W_1(0)\|_F^2}$ during training to verify Theorem 5.1 for two-layer linear and ReLU networks.

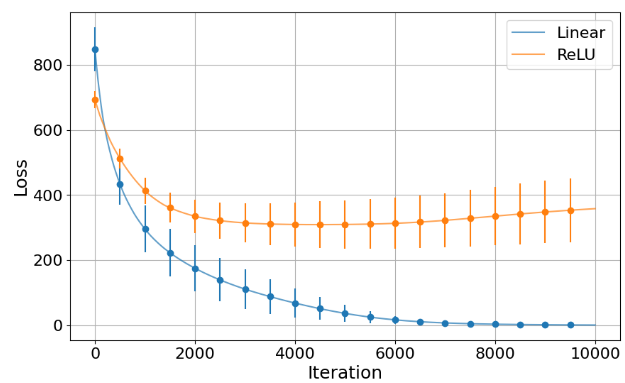

Figure 2. We plot the loss during training for two-layer linear and ReLU networks.

Launay et al., 2020). Our first main result concerns a conservation law for the FA learning dynamics that can explain the layer-wise alignment phenomena. Specifically, we introduce a conserved quantity, which remains invariant throughout the training process. This invariance holds under some general conditions related to the activation function and the loss function.

Theorem 5.1. *Suppose that we apply feedback alignment to a scalar output ReLU network with differentiable loss function. Then the flow of the layer weights under feedback alignment for all $t \in \mathbb{R}_{\geq 0}$ maintains,*

$$\frac{1}{2}\|W_i(t)\|_F^2 - \langle W_{i+1}(t), B_{i+1}\rangle$$
$$= \frac{1}{2}\|W_i(0)\|_F^2 - \langle W_{i+1}(0), B_{i+1}\rangle$$

Proof Sketch. The major technical part of the proof is to show that we have:

$$\langle \dot{W}_i, W_i\rangle = \langle \dot{W}_{i+1}, B_{i+1}\rangle.$$

The trace map is linear so after an integration by parts we have that,

$$\Rightarrow \int_0^t \mathrm{Tr}(\dot{W}_i W_i^T) ds = \mathrm{Tr}\left[\int_0^t \dot{W}_i W_i^T\right] ds$$
$$= \frac{1}{2}\|W_i(t)\|_F^2 - \frac{1}{2}\|W_i(0)\|_F^2$$

Finally,

$$\int_0^t \mathrm{Tr}(\dot{W}_{i+1} B_{i+1}^T) ds = \mathrm{Tr}\left[\int_0^t \dot{W}_{i+1} B_{i+1}^T ds\right]$$
$$= \langle W_{i+1}(t), B_{i+1}\rangle - \langle W_{i+1}(0), B_{i+1}\rangle$$

The result follows. □

This conservation law has several intriguing implications. One such implication is the existence of an implicit bias in FA that is analogous to, but distinct from, the bias in gradient descent. This bias effectively governs the learning trajectory of the FA rule, directing it towards certain types of solutions over others.

The conservation law also offers an insight into the alignment challenge in FA. As a corollary, if we initialize $W_{i+1}(0) = B_{i+1}$ such that $\|W_i(0)\| \leq \|W_{i+1}(0)\|$ then the conservation law simplifies to,

$$\langle W_{i+1}(t), B_{i+1}\rangle \geq \|W_{i+1}(0)\|_F^2 - \frac{1}{2}\|W_i(0)\|_F^2 \geq 0$$

Thus, there are general initialization schemes that guarantee layer-wise alignment.

Lastly, we focus on the case of two-layer linear networks trained with FA. By exploiting the conservation law, we show that these networks are capable of converging to a global optimum.

Theorem 5.2. *Assume that we are to fit data y with squared-loss and an over-parameterized two-layer network $f_{w_t}(X) = Xw_t = XW_1(t)W_2(t)$ with data X such that rows are linearly-independent. Assume we may pick $w_0 \in span(X^T)$ such that we have positive alignment for all time. If we run (direct) feedback alignment flow then we have the following,*

$$\lim_{t\to\infty} e^{rt} \cdot \|y - Xw_t\|_2 \to 0$$

for some $r > 0$. Moreover, $w_\infty = W_1(\infty)W_2(\infty)$ is the minimum-norm solution.

Proof Sketch. We prove an auxiliary lemma that shows that the network weights stay in the span of the input data matrix

X^T. To show this we derive the update rules in continuous time, under the assumption that the weights are intitialized in the span of the data. A simple option is to initialize with $W_1 = 0$. Following this, we show that the evolution of the error is decreasing geometrically over time implying that the algorithm will converge. Finally, we demonstrate that under these conditions, the solution found by the flow will be related to the Moore-Penrose psuedoinverse which minimizes the norm of the solution. □

This result implies that feedback-alignment applied to linear networks in the over-parameterized regime enjoy similar generalization properties to those of networks trained with gradient descent. These types of results have been popular as a way to explain why deep neural networks are able to generalize (Belkin et al., 2019).

These theoretical findings indicate that FA shares some properties with the more traditional gradient descent approach to learning. In particular, we also establish a connection between initialization, layer-wise alignment, and convergence with our results. This could set the stage for further exploration of alternative learning rules that can overcome the practical challenges associated with BP, whilst retaining its effectiveness.

6. Experiments

In our effort to comprehend the fundamental principles governing Feedback Alignment (FA), we devised a series of experiments to validate our theoretical assertions. We specifically aimed to observe the FA learning dynamics, understanding its implicit bias and ultimately, its convergence to the global optimum in over-parameterized two-layer linear networks.

Our network design involved multi-layer configurations, with the task of finding the global minimizers for the squared-loss function. For the generation of training data, we relied on a multivariate normal distribution, where the true response variable was a noiseless linear transformation of the input.

A crucial element of our experiments was implementing FA to update the network weights. Theorem 5.1 predicts that the quantity $\frac{1}{2}\|W_i(t)\|_F^2 - \langle W_{i+1}(t), B_{i+1}\rangle$ should remain invariant under the FA learning flow. Tracking this quantity provided empirical validation of this theoretical result. Furthermore, comparing the FA-learned weights to the Moore-Penrose pseudoinverse solution allowed us to observe the convergence of FA to the global optimum.

We carried out these procedures for two network architectures - a two-layer linear network and a two-layer ReLU network. To assure high-probability positive layer-wise alignment and convergence, we initialized the first layer weights using a distribution $\mathcal{N}(0, 1/10)$, and the second layer weights with a distribution $\mathcal{N}(0, 1)$. We initialized the feedback matrices to match the initialization of the second layer weights.

6.1. Discussion of Results

The experimental results, as displayed in Figures 1 and 2, corroborate the conservation law postulated by Theorem 5.1. The tracked quantity $\frac{1}{2}\|W_i(t)\|_F^2 - \langle W_{i+1}(t), B_{i+1}\rangle$ remained nearly constant across iterations which provides empirical support to our theoretical results. This manifestation of the conservation law substantiates the intriguing parallels between FA and Gradient Descent (GD), specifically with regards to their implicit biases. In the context of the two-layer linear network, the convergence of FA-learned weights to the global optimal solution was observed. We verified this convergence is identical to the minimum norm solution up to three significant digits. This underscores the implicit bias of feedback alignment towards solutions that generalize well.

In particular, we find the (nearly) exact conservation of layer-wise in non-linear networks updating with the feedback alignment learning rule to be compelling. While we do not show layer-wise alignment is sufficient for convergence in the non-linear setting, it seems plausible that such a condition would become useful in the large-width setting which has been successfully analyzed for gradient descent. Overall, the implications of these shared similarities warrant further investigation, potentially fostering the development of more powerful theoretical techniques capable of distinguishing these two learning rules. We think that research into establishing (or failing) implicit bias results for bio-plausible learning rules could elucidate higher-level principles behind good learning rules.

7. Conclusion

In conclusion, our findings challenge the prevailing narrative that FA and GD are fundamentally different learning algorithms. We demonstrate that, under certain conditions, FA can mirror the behavior of GD, offering a computationally efficient and biologically plausible alternative. Our main motivation is to pave the way for developing more efficient deep learning algorithms, better approximating the learning dynamics in biological systems. Overall, our results connecting layer-wise alignment with convergence in linear-models suggest that layer-wise alignment may also be a useful tool for analyzing learning dynamics in wide neural networks. Future work aims to extend these results to non-linear models, facilitating the creation of more biologically plausible models.

Acknowledgements

We'd like to thank Tengyu Ma for stimulating conversation about the project and inspiring the focus on investigating implicit bias presented.

References

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Bordelon, B. and Pehlevan, C. The influence of learning rule on representation dynamics in wide neural networks. *arXiv preprint arXiv:2210.02157*, 2022.

Du, S. S., Hu, W., and Lee, J. D. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. *Advances in neural information processing systems*, 31, 2018.

Duin, R. P. Classifiers in almost empty spaces. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, volume 2, pp. 1–7. IEEE, 2000.

Launay, J., Poli, I., Boniface, F., and Krzakala, F. Direct feedback alignment scales to modern deep learning tasks and architectures. *Advances in neural information processing systems*, 33:9346–9360, 2020.

Lechner, M. Learning representations for binary-classification without backpropagation. In *8th International Conference on Learning Representations*, 2020.

Lillicrap, T. P., Cownden, D., Tweed, D. B., and Akerman, C. J. Random synaptic feedback weights support error backpropagation for deep learning. *Nature communications*, 7(1):13276, 2016.

Nøkland, A. Direct feedback alignment provides learning in deep neural networks. *Advances in neural information processing systems*, 29, 2016.

Refinetti, M., d'Ascoli, S., Ohana, R., and Goldt, S. Align, then memorise: the dynamics of learning with feedback alignment. In *International Conference on Machine Learning*, pp. 8925–8935. PMLR, 2021.

Song, G., Xu, R., and Lafferty, J. Convergence and alignment of gradient descent with random backpropagation weights. *Advances in Neural Information Processing Systems*, 34:19888–19898, 2021.

Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

Stork, D. G. Is backpropagation biologically plausible. In *International Joint Conference on Neural Networks*, volume 2, pp. 241–246. IEEE Washington, DC, 1989.

Vallet, F., Cailton, J.-G., and Refregier, P. Linear and nonlinear extension of the pseudo-inverse solution for learning boolean functions. *Europhysics Letters*, 9(4):315, 1989.

A. Omitted Proofs

A.1. Proof of theorem 5.1

Theorem 5.1. *Suppose that we apply feedback alignment to a scalar output ReLU network with differentiable loss function. Then the flow of the layer weights under feedback alignment for all $t \in \mathbb{R}_{\geq 0}$ maintains,*

$$\frac{1}{2}\|W_i(t)\|_F^2 - \langle W_{i+1}(t), B_{i+1}\rangle$$

$$= \frac{1}{2}\|W_i(0)\|_F^2 - \langle W_{i+1}(0), B_{i+1}\rangle$$

Let A_i be a diagonal matrix with activations of the i-th output layer on the diagonal,

$$A_i = \text{diag} \circ \sigma' \circ W_i \circ \sigma \circ \ldots \circ \sigma \circ W_1(x)$$

This matrix indicates if an output at the i-th layer is non-zero. Supressing composition we have,

$$f_W(x) = xW_1A_1 \ldots A_{L-1}W_L$$

First we calculate the gradient and then compare with the feedback alignment update rule.

$$\Rightarrow \nabla_{W_i} f_W(x) = (xW_1A_1 \ldots W_{i-1}A_{i-1})^T(A_iW_{i+1} \ldots A_{L-1}W_L)^T$$

Feedback aligment simplifies the backward pass by replacing terms with random feedback matrices. For a (leaky) ReLU layer network learning with a differentiable loss such as $\mathcal{L} : \mathbb{R}^m \to \mathbb{R}^m \to \mathbb{R}_{\geq 0}$ where we take $\delta_L(t) = \nabla_{a_L}\mathcal{L}$ as the gradient at time t. We have the following learning dynamics for the i^{th} and $(i+1)^{\text{th}}$ layers under feedback alignment,

$$\dot{W}_i = -\eta \cdot \mathbb{E}\left[(xW_1A_1 \ldots W_{i-1}A_{i-1})^T \delta_L(t)^T (A_iB_{i+1} \ldots A_{L-1}B_L)^T\right]$$

$$\dot{W}_{i+1} = -\eta \cdot \mathbb{E}\left[(xW_1A_1 \ldots W_iA_i)^T \delta_L(t)^T (A_{i+1}B_{i+2} \ldots A_{L-1}B_L)^T\right]$$

where $B_i \in \mathbb{R}^{d_i \times d_{i+1}}$ is a random feedback operator fixed at initialization.

For the first-layer we have,

$$\langle \dot{W}_i, W_i\rangle = \text{Tr}(\dot{W}_i^T W_i)$$

$$= -\eta \cdot \mathbb{E}_k\left[\text{Tr}((A_iB_{i+1} \ldots A_{L-1}B_L)\delta_L(t)(x_kW_1A_1 \ldots W_{i-1}A_{i-1})W_i)\right]$$

$$= -\eta \cdot \mathbb{E}_k\left[\phi_k'(t) \cdot \text{Tr}(x_kW_1A_1 \ldots W_iA_iB_{i+1} \ldots B_L)\delta_L(t)\right]$$

We are making use of the trace representation for the inner-product and the cyclic property of the trace map. Similarly, for the next layer we obtain,

$$\langle \dot{W}_{i+1}, B_{i+1}\rangle = \text{Tr}(\dot{W}_{i+1}^T B_{i+1})$$

$$= -\eta \cdot \mathbb{E}_k\left[\text{Tr}((A_{i+1}B_{i+2} \ldots A_{L-1}B_L)\delta_L(t)(x_kW_1A_1 \ldots W_iA_i)B_{i+1})\right]$$

$$= -\eta \cdot \mathbb{E}_k\left[\text{Tr}(x_kW_1A_1 \ldots W_iA_iB_{i+1} \ldots B_L)\delta_L(t)^T\right]$$

The major implication is that,

$$\langle \dot{W}_i, W_i\rangle = \langle \dot{W}_{i+1}, B_{i+1}\rangle$$

The trace map is linear so after an integration by parts we have that,

$$\Rightarrow \int_0^t \text{Tr}(\dot{W}_i W_i^T)ds = \text{Tr}\left[\int_0^t \dot{W}_i W_i^T\right] ds$$

$$= \frac{1}{2}\|W_i(t)\|_F^2 - \frac{1}{2}\|W_i(0)\|_F^2$$

Finally,

$$\int_0^t \text{Tr}(\dot{W}_{i+1} B_{i+1}^T) ds = \text{Tr}\left[\int_0^t \dot{W}_{i+1} B_{i+1}^T ds\right]$$
$$= \langle W_{i+1}(t), B_{i+1}\rangle - \langle W_{i+1}(0), B_{i+1}\rangle$$

The result follows. □

Remark: if we initialize $W_i(0) = W_{i+1}(0) = 0$ then the conservation law simplifies to,

$$\langle W_{i+1}(t), B_{i+1}\rangle = \frac{1}{2}\|W_i(t)\|_F^2 \geq 0$$

Thus, feedback alignment with the zero-initialization preserves non-negative alignment.

A.2. Proof of Theorem 5.2

There is another invariant that is useful for analyzing the type of solution we obtain from the flow of (direct) feedback alignment.

Lemma A.1. *If we have $w(t) = W_1(t)W_2(t) \in$ span(X^T) for $t = 0$ then have $w(t) \in$ span(X^T) for all time.*

Proof. Under flow the first-layer parameters update according to,

$$\dot{W}_1(t) = -\eta X^T e(t) B^T$$

$$W_1(0) \in \text{span}(X^T)$$

Observe that $\dot{W}_1$ is invariantly in the span of X^T so we may conclude that $W_1(t)$ is always in the span of X^T. This means we have $W_1(t) = X^T\alpha_0(t)$ for all time. Extending to $W = W_1W_2$,

$$\dot{W}(t) = \dot{W}_1 W_2 + W_1 \dot{W}_2$$

$$= -\eta(X^T e B^T W_2 + W_1 W_1^T X^T e)$$
$$= -\eta(X^T\alpha_1(t) + X^T\alpha_2(t)) \in \text{span}(X^T)$$

We can calculate an iteration for the error vector as,

$$\Rightarrow e(t+1) = XW_1(t+1)W_2(t+1) - y$$

$$= X\left(W_1(t) - \eta X^T e(t) B^T\right)\left(W_2(t) - \eta W_1(t)^T X^T e(t)\right) - y$$
$$= X\left(W_1(t)W_2(t) - \eta(X^T e(t) B^T W_2(t) + W_1(t)W_1(t)^T X^T e(t)) + \eta^2 X^T e(t) B^T W_1(t)^T X^T e(t)\right) - y$$
$$= e(t) - \eta(XX^T e(t) B^T W_2(t) + XW_1(t)W_1(t)^T X^T e(t)) + \eta^2\left(XX^T e(t) B^T W_1(t)^T X^T e(t)\right)$$
$$= \left(I - \eta((B^T W_2(t))XX^T + XW_1(t)W_1(t)^T X^T)\right)e(t) + \eta^2\left(XX^T e(t) B^T W_1(t)^T X^T e(t)\right)$$
$$\simeq (I - \eta F)e(t)$$

In the last step we observe that $B^T W_2(t)$ is an inner-product. We can fix this quantity positive for all time using proper initialization and Lemma 1 and so we have $F \succ 0$. Taking the limit $\eta \to 0$ yields the flow,

$$\dot{e} = -F(t)e$$

which will decrease to zero at a geometric rate.

So under the flow we'll suppose the parameters update according to,

$$\dot{w} = -\eta X^T e$$

$$w(0) = 0$$

Observe that $\dot{w}$ is invariantly in the span of X^T so we may conclude that $w(t)$ is always in the span of X^T. Generically, any solution in the over-parameterized setting is a global optimizer such that $Xw = y$. This means that the limit of the flow can be written as $w^* = X^T\alpha$ for some coefficient vector with the constraint that $Xw^* = y$. After some manipulations we find that,

$$\begin{aligned} y = Xw^* = XX^T\alpha \\ \Rightarrow \alpha = (XX^T)^{-1}y \\ \Rightarrow w^* = X^T(XX^T)^{-1}y = X^+y \end{aligned}$$

This means that the solution X^+ picked from gradient flow is the Moore-Penrose psuedoinverse. This can be defined as the matrix,

$$X^+ = \lim_{\lambda \to 0^+} X^T(XX^T + \lambda I)^{-1}$$

Also observe that there is a unique minimizer for the regularized problem,

$$\min_w L(w) + \lambda \|w\|_2^2$$

with value $w_\lambda = X^T(XX^T + \lambda I)^{-1}y$. Perhaps, $Xw = y$ has a set of solutions, but it is clear this set is convex so there is a unique minimum norm solution. On the other hand, each w_λ corresponds to a best solution with norm below the minimum. However, we have $w^* = \lim_{\lambda \to 0^+} w_\lambda$ from continuity. Since w^* is an exact solution it can't have less than the minimum-norm and it is clear w^* can't have above the minimum-norm either since this is not the case for any of the w_λ. We conclude that gradient descent does indeed find the minimum norm solution.

$\square$