Solving Inverse Problems with Deep Linear Neural Networks: Global Convergence Guarantees for Gradient Descent with Weight Decay

Hannah Laus*[1,2], Suzanna Parkinson[3], Vasileios Charisopoulos[4], Felix Krahmer[1,2,5], and Rebecca Willett[3,4,6,7,8,9]

[1]Department of Mathematics, Technical University of Munich
[2]Munich Center for Machine Learning (MCML)
[3]Committee on Computational and Applied Mathematics, University of Chicago
[4]Data Science Institute, University of Chicago
[5]Munich Data Science Institute (MDSI), Technical University of Munich
[6]Department of Computer Science, University of Chicago
[7]Department of Statistics, University of Chicago
[8]NSF-Simons National Institute for Theory and Mathematics in Biology (NITMB)
[9]NSF-Simons National Institute for AI in the Sky (SkAI)

Abstract

Machine learning methods are commonly used to solve inverse problems, wherein an unknown signal must be estimated from few measurements generated via a known acquisition procedure. In particular, neural networks perform well empirically but have limited theoretical guarantees. In this work, we study an underdetermined linear inverse problem that admits several possible solution mappings. A standard remedy (e.g., in compressed sensing) establishing uniqueness of the solution mapping is to assume knowledge of latent low-dimensional structure in the source signal. We ask the following question: do deep neural networks adapt to this low-dimensional structure when trained by gradient descent with weight decay regularization? We prove that mildly overparameterized deep linear networks trained in this manner converge to an approximate solution that accurately solves the inverse problem while implicitly encoding latent subspace structure. To our knowledge, this is the first result to rigorously show that deep linear networks trained with weight decay automatically adapt to latent subspace structure in the data under practical stepsize and weight initialization schemes. Our work highlights that regularization and overparameterization improve generalization, while overparameterization also accelerates convergence during training.

1 Introduction

Machine learning approaches, especially those based on deep neural networks, have risen to prominence for solving a broad class of inverse problems. In particular, deep learning approaches constitute the state of the art for various inverse problems arising in medical imaging (e.g. MRI or CT) [1, 2, 3, 4], image denoising [1, 5], and image inpainting [6, 7]. Despite its impressive performance for inverse problems, almost all the theoretical underpinnings of deep learning focus

*Corresponding author. Email: hannah.laus@tum.de

on regression or classification problems; see [8] for a summary of the theoretical results for deep neural networks for inverse problems. On the other hand, there is a strong need for theory: understanding the behavior of deep neural networks is crucial when they are deployed in critical applications such as medical imaging.

A challenge is that neural networks are typically trained on a subset of all potential data points – pairs of "realistic" signals and their measurements – the distribution of which is not known a priori. Nevertheless, one aims for robustness: perturbed measurements should yield approximate reconstructions, even if the perturbation no longer corresponds to a realistic signal passed through the forward model. As suggested by a number of works [9, 10, 11, 12], the robustness of machine learning approaches is by no means automatic and requires special attention. Even for the most fundamental model of a set of signals lying in a subspace, this effect is observed in numerical simulations. For example, Figure 1a shows that a linear network trained via vanilla gradient descent (i.e., with zero regularization, $\lambda = 0$) on synthetic data starting from a random initialization converges to a solver that is not very robust to perturbations (here, Gaussian noise). In Figure 1b we can see the same effect for a non-linear ReLU network trained on data from a union of subspaces model; see Appendix C.

In this paper, we discuss ways out of this fundamental bottleneck focusing, as a proof of concept, on the aforementioned model of a high-dimensional signal lying in an (unknown) low-dimensional subspace. Indeed, Figure 1 shows that robustness considerably improves in the presence of ℓ_2-regularization (also known as *weight decay*), a standard strategy in machine learning designed to promote simple parameter configurations [13, 14]. For the purposes of analysis, we address the case where the training data corresponds to solved linear inverse problems:

$$\begin{aligned} X &= \begin{bmatrix} x^1 & \dots & x^n \end{bmatrix} \in \mathbb{R}^{d \times n}, && \text{(signals)} \\ Y &= \begin{bmatrix} y^1 & \dots & y^n \end{bmatrix} \in \mathbb{R}^{m \times n}, && \text{(measurements)} \\ \text{where} \quad y^i &= Ax^i, \;\; x^i \in \operatorname{range}(R) \quad \text{for all } i. && (1) \end{aligned}$$

Here, $A \in \mathbb{R}^{m \times d}$ is a fixed measurement operator with $m < d$ and $R \in \mathbb{R}^{d \times s}$ is a (unknown) matrix with orthogonal columns that span a low-dimensional subspace (i.e., $s \ll d$). One aims to solve the regularized minimization problem

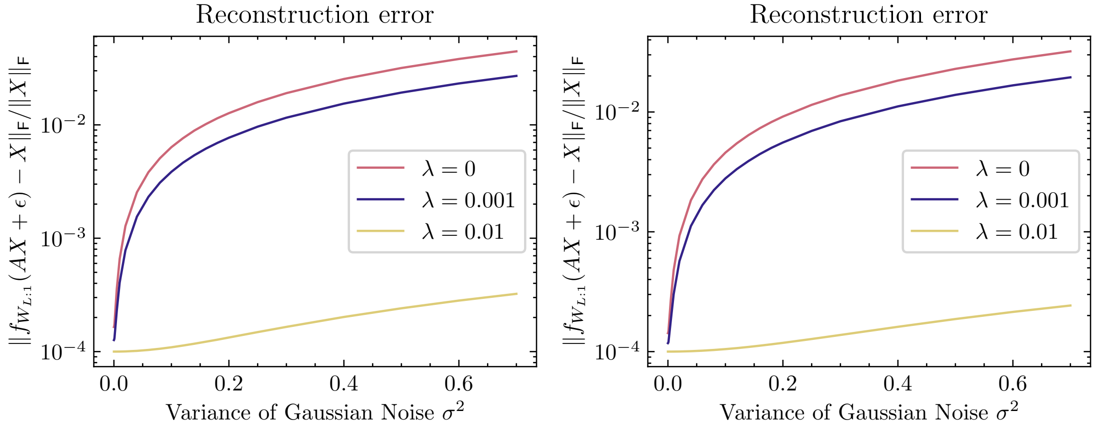
$$\min_{W_1,\dots,W_L} \|f_{W_{L:1}}(Y) - X\|_{\mathsf{F}}^2 + \lambda \sum_{\ell=1}^{L} \|W_\ell\|_{\mathsf{F}}^2 = \min_{W_1,\dots,W_L} \sum_{i=1}^{n} \|f_{W_{L:1}}(y^i) - x^i\|^2 + \lambda \sum_{\ell=1}^{L} \|W_\ell\|_{\mathsf{F}}^2. \quad (2)$$

Here, $f_{W_{L:1}}$ is a depth-L neural network with weight matrices $W_1, \dots, W_L$ for some $L \in \mathbb{N}$. It is not hard to show that, for small λ, the global minimizer of this non-convex problem yields a robust solution – namely, it has the following two properties:

1. It is accurate (with error vanishing in the limit as $\lambda \to 0$) on the image of the signal subspace – that is, it accurately reconstructs signals from their measurements.

2. It is zero on the orthogonal complement of the image of the signal subspace – that is, perturbations orthogonal to the model are eliminated (see Lemma B.1).

The remaining issue is that, due to the non-convexity of the problem (2), no algorithms with global convergence guarantees are available to date (to the best of our knowledge). As a proxy, practitioners typically apply gradient descent or stochastic gradient descent to the regularized



(a) Linear network **(b)** ReLU network with extra linear layers

Figure 1: Comparison of robustness against Gaussian noise for training with and without weight decay. All experiments use signals of dimension $d = 200$ and measurements of dimension $m = 100$; all networks have $L = 5$ layers and hidden layer width $d_w = 400$. In Figure 1a, the model is a linear neural network trained on data lying in a subspace of dimension $s = 5$. In Figure 1b, the model is a ReLU neural network with extra linear layers, similar to the setup of Parkinson et al. [15], trained on data that lie in the union of three subspaces, each of dimension $s = 5$. A detailed description of the numerical experiment can be found in Section 3 and Appendix C.

objective. However, whether this approach produces a good approximation to the desired global minimizer (or *any* point that shares the aforementioned properties) remains unclear.

In this paper, we provide an answer for fully connected deep linear neural networks $f_{W_{L:1}}(Y) = W_L \cdots W_1 Y$ trained by gradient descent on the regularized objective (2). Our contributions can be summarized as follows (see Theorem 2.3).

1. We show that gradient descent converges to an approximate solution that reconstructs signals from their measurements with error vanishing in the limit as $\lambda \to 0$.

2. We show that the part of the weights acting on the orthogonal complement of the image of the signal subspace is small after a finite number of iterations.

3. We show that optimizing the regularized objective (2) leads to a more robust solution than in the non-regularized case (see Section 2.2).

1.1 Related work

Benefits of weight decay for generalization. It is believed that for understanding generalization properties of neural networks, "the size of the weights is more important than the size of the network" [16]. This idea has been studied in several works [17, 18, 19, 20, 21, 22], and is especially notable in light of modern machine learning that operates in highly overparameterized regimes [23]. Regularizing the ℓ_2-norm of the parameters (i.e., weight decay) to encourage small-norm weight matrices is common practice in neural network training and has been empirically observed to improve generalization [13, 14, 24, 25].

Multiple works have addressed the properties of global minimizers of the ℓ_2-regularized loss and of minimal-norm interpolants of the data [26, 27, 28, 29, 30]. Several works have found that such networks adapt to low-dimensional structure [31, 15, 32, 33]. In particular, minimal-norm linear deep neural networks are known to induce low-rank mappings [34, 35].

Convergence of gradient descent for deep linear networks. Several works study the dynamics of gradient descent for training deep linear neural networks in general regression tasks under different assumptions. For example, Du and Hu [36] show that gradient descent starting from a random Gaussian initialization will converge at a linear rate to a global minimizer of the *unregularized* loss ($\lambda = 0$) as long as the hidden layer width scales linearly in the input dimension and depth; a closely related work by Hu et al. [37] demonstrates that the hidden layer width no longer needs to scale with the network depth when weights are initialized according to an orthogonal scheme. Similarly, Arora et al. [38] study convergence of gradient descent when (i) weight matrices at initialization are approximately balanced and (ii) the problem instance satisfies a "deficiency margin" property ruling out certain rank-deficient solutions – a condition later removed by the analysis of Nguegnang et al. [39]. On the other hand, Xu et al. [40] show that gradient descent converges to a global minimum for linear neural networks with two layers and mild overparameterization *without* any assumptions on the initialization; however, their proof does not readily extend to neural networks of arbitrary depth L. Shamir [41] studies gradient descent on deep linear networks when the dimension and hidden width are both equal to one. Other results include Kawaguchi [42] and Laurent and von Brecht [43], who show that under certain assumptions, all local minima are global. Finally, a number of works focus on gradient flow [44, 45, 46, 47, 48, 49], the continuous-time analog of gradient descent.

All the works mentioned so far study gradient descent or gradient flow without any explicit regularization. In contrast, Arora et al. [50] study the ℓ_2-regularized objective for deep linear networks but do not focus on the effects of the regularization in the analysis. Instead, they show that depth has a preconditioning effect that accelerates convergence. However, their analysis for the discrete-time setting relies on near-zero initialization and small stepsizes. Lewkowycz and Gur-Ari [51] study the regularization effect of weight decay for *infinitely wide* neural networks with positively homogeneous activations, finding that model performance peaks at approximately λ^{-1} iterations – a finding also supported by our analysis (cf. Theorem 2.3). However, their theoretical analysis only covers gradient flow updates. The works [52, 53], inspired by the LoRA technique [54], show that gradient descent updates of deep linear networks traverse a "small" subspace when the input data lies on a low-dimensional structure. Unfortunately, their proofs (i) rely on an "orthogonal initialization" scheme and (ii) do not provide any guarantees on the accuracy of the solution learned by gradient descent. Finally, Wang and Jacot [55] study the implicit bias of (stochastic) gradient descent for deep linear networks. They show that SGD with sufficiently small weight decay initially converges to a solution that overestimates the rank of the true solution mapping, but SGD will find a low-rank solution with positive probability given a sufficiently large number of epochs (proportional to $O(\eta^{-1}\lambda^{-1})$). However, their work does not rule out the possibility that the low-rank solution found by SGD is a poor fit to the data.

1.2 Notation and basic constructions

We briefly introduce the notation used in the paper. We write $\|A\|_{\mathsf{F}} := \sqrt{\mathsf{Tr}(A^\mathsf{T} A)}$ for the *Frobenius norm* of a matrix $A \in \mathbb{R}^{m \times d}$ and $\|A\|_{\mathsf{op}} := \sup_{x : \|x\|=1} \|Ax\|$ for its *spectral norm*. Moreover, we let $A^\dagger$ denote the *Moore-Penrose* pseudoinverse of A. We write $\sigma_{\min}(A)$ for the smallest *nonzero* singular value of A. The vectorization operator **vec** transforms a matrix $A \in \mathbb{R}^{m \times d}$ into a vector $\mathbf{vec}(A) \in \mathbb{R}^{md}$ in column-major order. We let $A \otimes B$ denote the

Algorithm 1 Gradient descent

Input: data X, Y, step-size $\eta > 0$, iterations T.
Initialize weights $\{W_\ell(0)\}_{\ell=1}^{L}$.
for $t = 0, 1, \dots, T-1$ **do**
  $W_\ell(t+1) = W_\ell(t) - \eta \nabla \mathcal{L}(\{W_\ell(t)\}_{\ell=1}^{L}; (X, Y))$
end for
return $\{W_\ell(T)\}_{\ell=1}^{L}$.

Kronecker product between matrices A and B; for compatible A, X and B, the Kronecker product and **vec** operator satisfy

$$\mathbf{vec}(AXB^\mathsf{T}) = (B \otimes A) \cdot \mathbf{vec}(X). \tag{3}$$

Given a projection matrix P (i.e., a symmetric, idempotent matrix), we write $P_\perp := I - P$ for the projection matrix onto the orthogonal complement of range(P). Finally, given scalars A and B, we write $A \lesssim B$ to indicate that there is a dimension-independent constant $c > 0$ such that $A \leq cB$; the precise value of c may change between occurrences.

2 Main result

In this section, we present our main result as well as a proof sketch focusing on the depth $L = 2$ case. Recall that we are interested in solving (2), for the special case where $f_{W_{L:1}}$ is a deep linear network, using gradient descent (Algorithm 1). Concretely, we want to minimize the following loss function:

$$\mathcal{L}(\{W_\ell\}_{\ell=1,\dots,L}; (X, Y)) := \frac{1}{2}\|W_L \cdots W_1 Y - X\|_\mathsf{F}^2 + \frac{\lambda}{2}\sum_{\ell=1}^{L}\|W_\ell\|_\mathsf{F}^2. \tag{4}$$

We consider weight matrices of the following sizes:

- The weight matrix of the input layer $W_1 \in \mathbb{R}^{d_w \times m}$, where d_w is a width common to all hidden layers.
- The weight matrix of the output layer $W_L \in \mathbb{R}^{d \times d_w}$.
- All other weight matrices $W_2, \dots, W_{L-1} \in \mathbb{R}^{d_w \times d_w}$.

We will also write $W_{j:i}(t)$ for the following product of weight matrices at the t^{th} iteration:

$$W_{j:i}(t) := \prod_{\ell=j}^{i} W_\ell(t). \tag{5}$$

Having fixed the architecture, we introduce two mild assumptions under which our results hold.

Assumption 2.1 (Restricted Isometry Property)**.** The measurement matrix A from (1) satisfies the following: there exists $\delta > 0$ such that, for all vectors $x \in \text{range}(R)$,

$$(1-\delta)\|x\|^2 \leq \|Ax\|^2 \leq (1+\delta)\|x\|^2. \tag{6}$$

Assumption 2.1 is standard in the compressed sensing literature [56], as it is a sufficient condition that enables the solution of high-dimensional linear inverse problems from few measurements. In our context, Assumption 2.1 essentially states that the training data has been sampled from inverse problems that are identifiable.

Our next assumption relates to the network initialization:

Assumption 2.2 (Initialization)**.** The weight matrices $W_1, \dots, W_L$ at initialization are sampled from a scaled ("fan-in") normal distribution:

$$[W_\ell(0)]_{ij} \overset{\text{i.i.d.}}{\sim} \begin{cases} \mathcal{N}\left(0, \frac{1}{m}\right), & \ell = 1, \\ \mathcal{N}(0, \frac{1}{d_w}), & \ell = 2, \dots, L. \end{cases} \tag{7}$$

Assumption 2.2 is by no means restrictive: it was introduced by [57] as a heuristic for stabilizing neural network training and enjoys widespread adoption.[1]

We now present an informal version of our main result. The formal statement can be found in Theorem A.2, and the proof comprises Appendices A.1 to A.7.

Theorem 2.3 (Informal)**.** *Let Assumptions 2.1 and 2.2 hold and set the step size η and weight decay parameter λ as*

$$\eta := {}^{m}/_{L \cdot \sigma^2_{\max}(X)}, \quad \lambda := \gamma \sigma^2_{\min}(X) \sqrt{{}^{m}/_{d}}, \tag{8}$$

where $\gamma \in (0, 1]$ is a user-specified accuracy parameter. Moreover, define the times

$$\tau = \inf\left\{ t \in \mathbb{N} \mid \|W_{L:1}(t)Y - X\|_{\mathsf{F}} \le \frac{80\gamma\|X\|_{\mathsf{F}}}{L} \right\}, \tag{9a}$$

$$T = \frac{2L\kappa^2 \log(d_w)}{\gamma} \sqrt{\frac{d}{m}}, \tag{9b}$$

where $\kappa := \|X\|_{\mathsf{op}}\|X^\dagger\|_{\mathsf{op}}$ denotes the condition number of X. Finally, let $\mathrm{sr}(X) := {}^{\|X\|^2_{\mathsf{F}}}/_{\|X\|^2_{\mathsf{op}}}$ denote the stable rank *of X. Then, as long as the hidden layer width satisfies*

$$d_w \gtrsim d \cdot \mathrm{sr}(X) \cdot \mathrm{poly}(L, \kappa),$$

gradient descent (Algorithm 1) produces iterates that satisfy

$$\|W_{L:1}(t+1)Y - X\|_{\mathsf{F}} \le \begin{cases} \left(1 - \frac{1}{32\kappa^2}\right) \|W_{L:1}(t)Y - X\|_{\mathsf{F}}, & t < \tau; \\ C_1\gamma\|X\|_{\mathsf{F}}, & \tau \le t \le T \end{cases} \tag{10}$$

$$\|W_{L:1}(T)P^{\perp}_{\mathrm{range}(Y)}\|_{\mathsf{op}} \le \left(\frac{1}{d_w}\right)^{C_2}, \tag{11}$$

where C_1 and C_2 are universal positive constants. These guarantees hold with high probability over the random initialization.

Equation (10) in Theorem 2.3 demonstrates that the reconstruction of X from Y can be made arbitrarily accurate using a suitably small choice of regularization parameter λ. On the other hand, Equation (11) ensures that the component of the weights acting on the orthogonal complement of the signal subspace can be made small by increasing the hidden width of the

[1] See, e.g., the `torch.nn.init.kaiming_normal_` initialization method in `Pytorch`.

model; this ensures robustness to noisy test data as discussed below in Section 2.2. Theorem 2.3 also highlights two distinct phases of gradient descent; during the first τ iterations, Equation (10) suggests that the error in the reconstruction converges linearly up to the threshold specified in (9a). Upon reaching that threshold, the behavior changes: while the reconstruction error can increase mildly from iteration τ to T, the component of the weights acting on the orthogonal complement of the signal subspace shrinks to the level shown in Equation (11). The number of iterations T of gradient descent required to achieve this behavior grows only logarithmically with the hidden width, but is highly sensitive to the targeted reconstruction accuracy – and therefore the weight decay parameter λ.

Remark 2.4. Plugging η and λ into Equation (9b) implies that $T = O(1/\eta\lambda)$; this is consistent with the results of Lewkowycz and Gur-Ari [51], Wang and Jacot [55]. The former work observes empirically that SGD without momentum attains maximum performance at roughly $O(1/\eta\lambda)$ iterations, while the latter work [55, Theorem B.2] suggests that stochastic gradient descent requires a similar number of iterations to find a low-rank solution — albeit one that might be a poor data fit.

Remark 2.5. Theorem 2.3 remains valid when the step size η is chosen to be smaller than the value specified in the theorem, albeit at the expense of an increased number of iterations T.

2.1 Proof sketch

In this section, we provide a proof sketch for Theorem 2.3; full proofs are deferred to the Appendix. For simplicity, the proof sketch focuses on the case

$$L = 2, \;\; \kappa = 1, \;\; \delta = \frac{1}{10}.$$

Since normalization at initialization (Assumption 2.2) is essential to the proof, it is convenient to be explicit about normalization factors. We consider the equivalent loss

$$\mathcal{L}(W_1, W_2) := \frac{1}{2}\left\| \frac{1}{\sqrt{d_w m}} W_2 W_1 Y - X \right\|_{\mathsf{F}}^2 + \frac{\lambda}{2}\left(\frac{\|W_1\|_{\mathsf{F}}^2}{m} + \frac{\|W_2\|_{\mathsf{F}}^2}{d_w} \right), \tag{12}$$

under the assumption that $(W_1(0))_{ij}, (W_2(0))_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. We will also use the shorthand notation

$$\Phi(t) := \frac{1}{\sqrt{d_w m}} W_{2:1}(t) Y - X.$$

We refer to $\|\Phi(t)\|_{\mathsf{F}}$ as the regression error. Note that the gradient descent updates lead to the following decomposition:

$$\begin{aligned} & W_{2:1}(t+1) \\ & = \left(1 - \frac{\eta\lambda}{d_w}\right)\left(1 - \frac{\eta\lambda}{m}\right) W_{2:1}(t) + E_0(t) \\ & \quad - \frac{\eta}{\sqrt{d_w m}}\left(1 - \frac{\eta\lambda}{d_w}\right) W_2(t)(W_2(t))^{\mathsf{T}} \Phi(t) Y^{\mathsf{T}} \\ & \quad - \frac{\eta}{\sqrt{d_w m}}\left(1 - \frac{\eta\lambda}{m}\right) \Phi(t) Y^{\mathsf{T}} (W_1(t))^{\mathsf{T}} W_1(t), \end{aligned}$$

where $E_0(t) \in O(\eta^2)$ contains high-order terms. Multiplying both sides from the right by $(1/\sqrt{d_w m})Y$, subtracting X and taking norms, we obtain the following bound on the regression error at time $t+1$:

$$\|\Phi(t+1)\|_{\mathsf{F}} \leq \|I - \eta P(t)\|_{\mathsf{op}} \|\Phi(t)\|_{\mathsf{F}} + O\left(\frac{\eta\lambda}{m}\right) \cdot \left\|\frac{W_{2:1}(t)Y}{\sqrt{d_w m}}\right\|_{\mathsf{F}} + \left\|\frac{1}{\sqrt{d_w m}} E_0(t) Y\right\|_{\mathsf{F}}, \tag{13}$$

where $P(t)$ is an operator acting on matrix space whose matrix representation in terms of the vectorization is given by

$$P(t) := \frac{1}{d_w m}\left(1 - \frac{\eta\lambda}{m}\right)(W_1(t)Y)^{\mathsf{T}}(W_1(t)Y) + \frac{1}{d_w m}\left(1 - \frac{\eta\lambda}{d_w}\right)(Y^{\mathsf{T}}Y) \otimes ((W_2(t))^{\mathsf{T}} W_2(t)).$$

In particular, one can show that the high-order terms from $E_0(t)$ can be "folded" into the first term in (13), since

$$\left\|\frac{1}{\sqrt{d_w m}} E_0(t) Y\right\|_{\mathsf{F}} \leq \frac{\eta \|P(t)\|_{\mathsf{op}}}{4} \|\Phi(t)\|_{\mathsf{F}}. \tag{14}$$

Consequently, it is the spectrum of $P(t)$ that controls the rate of convergence (up to error that vanishes as $\lambda \to 0$). Thanks to properties of the Kronecker product, controlling the spectrum of $P(t)$ can be reduced to controlling the extremal singular values of $W_1 Y$ and W_2 (see Lemma A.7 for the full statement).

The remainder of the proof outlines two phases for the convergence behavior of gradient descent. In the first phase, the regression error is driven rapidly to a level that depends on the regularization strength λ; in the second phase, the "off-subspace" error decreases while the regression error can fluctuate, albeit in a controlled manner.

Phase 1: Rapid linear convergence. In the first phase, we show that the following properties hold by induction for $t < \tau$:

- (**Singular value control**): For all i, it holds that

$$\frac{3}{4}\sqrt{d_w} \leq \sigma_i(W_2(t)) \leq \frac{5}{4}\sqrt{d_w}$$
$$\frac{3}{4}\sqrt{d_w}\sigma_{\min}(X) \leq \sigma_i(W_1(t)Y) \leq \frac{5}{4}\sqrt{d_w}\sigma_{\max}(X).$$

- (**Small displacement**): We have

$$\|W_1(t) - \left(1 - \frac{\eta\lambda}{m}\right)^t W_1(0)\|_{\mathsf{op}} \lesssim \sqrt{d\,\mathrm{sr}(X)};$$
$$\|W_2(t) - \left(1 - \frac{\eta\lambda}{d_w}\right)^t W_2(0)\|_{\mathsf{op}} \lesssim \sqrt{d\,\mathrm{sr}(X)}.$$

- (**Sufficient decrease in regression error**): We have

$$\begin{aligned} &\|\Phi(t+1)\|_{\mathsf{F}} \\ &\leq \left(1 - \frac{\eta\sigma_{\min}^2(X)}{8m}\right)\|\Phi(t)\|_{\mathsf{F}} + \frac{5\eta\lambda}{2m}\sqrt{\frac{d}{m}}\|X\|_{\mathsf{F}}. \end{aligned} \tag{15}$$

A detailed argument can be found in Appendix A.5.

While $t < \tau$, where τ is defined in Theorem 2.3, the second term in the right-hand side of (15) satisfies

$$\frac{5\eta\lambda}{2m}\sqrt{\frac{d}{m}}\|X\|_{\mathsf{F}} \leq \frac{\eta\sigma_{\min}^2(X)}{16m}\|\Phi(t)\|_{\mathsf{F}}.$$

Consequently, we obtain the following recurrence for $t < \tau$:

$$\|\Phi(t+1)\|_{\mathsf{F}} \leq \left(1 - \frac{\eta\sigma_{\min}^2(X)}{16m}\right)\|\Phi(t)\|_{\mathsf{F}}. \tag{16}$$

Plugging $\eta = m/2\sigma_{\max}^2(X)$ into (16) yields the bound (10) for $t < \tau$. Iterating (16) until the condition in the definition of τ fails reveals that the length of phase 1 is at most

$$\tau \lesssim \log\left(\frac{1}{\gamma}\sqrt{\frac{d}{m}}\right) \quad \text{iterations.} \tag{17}$$

Consequently, we achieve regression error $O(\gamma)$ within $O(\log\frac{1}{\gamma})$ iterations, a rate commensurate with that achieved by gradient descent when minimizing the *convex* objective $\min_W \|WY - X\|_{\mathsf{F}}^2$ [36]. Reducing the *off-subspace* error, $\|W_{2:1}(t)P_{\mathrm{range}(Y)}^{\perp}\|_{\mathsf{op}}$, requires further work as outlined below.

Phase 2: Off-subspace component reduction. During the first phase, the off-subspace error will generally decrease but remain nontrivial, requiring additional iterations to bring to acceptable levels. The challenge is that when $t > \tau$, the regression error $\|\Phi(t)\|_{\mathsf{F}}$ is no longer monotonic. To that end, we argue that the regression error *remains* small (up to a constant multiplicative factor) for the next $O(\frac{1}{\gamma})$ steps, subject to the same stepsize requirements; in turn, these steps are sufficient to reduce the off-subspace error to $O(\mathrm{poly}(d_w^{-1}))$. Specifically, we argue that the following properties hold (see Appendix A.6) for all iterations t satisfying $\tau \leq t \leq O(\log(d_w)/\lambda)$:

- (**Singular value control II**): For all i, it holds that

$$\frac{5}{7}\sqrt{d_w} \leq \sigma_i(W_2(t)) \leq \frac{9}{7}\sqrt{d_w};$$
$$\sigma_{\max}(W_1(t)Y) \leq \frac{9}{7}\sqrt{d_w}\sigma_{\max}(X).$$

- (**Small displacement II**): We have that

$$\|W_1(t) - \left(1 - \frac{\eta\lambda}{m}\right)^{t-\tau} W_1(\tau)\|_{\mathsf{op}} \lesssim \sqrt{d\,\mathrm{sr}(X)}\log(d_w);$$
$$\|W_2(t) - \left(1 - \frac{\eta\lambda}{d_w}\right)^{t-\tau} W_2(\tau)\|_{\mathsf{op}} \lesssim \sqrt{d\,\mathrm{sr}(X)}\log(d_w).$$

- (**Small regression error**): We have

$$\|\Phi(t)\|_{\mathsf{F}} \lesssim \frac{\lambda\|X\|_{\mathsf{F}}}{\sigma_{\min}^2(X)}\sqrt{\frac{d}{m}} = O(\gamma\|X\|_{\mathsf{F}}).$$

Equipped with the properties above, we show that the off-subspace error satisfies the bound:

$$\|W_{2:1}(t)P^{\perp}_{\mathrm{range}(Y)}\|_{\mathsf{op}} \lesssim \left(1 - \frac{\lambda}{2}\right)^t \sqrt{\frac{d_w}{m}}, \tag{18}$$

which is at most $d_w^{-C_2}$ when $t \geq \Omega\left(\frac{\log(d_w)}{\lambda}\right)$. See Appendix A.7 for details.

2.2 Robustness to noisy test data

Training a network on the regularized objective with gradient descent leads to a more robust solution than a network trained without regularization. The following Corollary, whose proof can be found in Appendix A.8, formalizes this by considering a test instance with noisy measurements.

Corollary 2.6. *Let $(W_1(T), \dots, W_L(T))$ be the weight matrices of a deep linear network trained for T iterations in the setting of Theorem 2.3. Consider a test data point (x, y) satisfying $y = Ax + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Then, with high probability, the output of the network $W_{L:1}(T)(y)$ satisfies*

$$\|W_{L:1}(T)y - x\| \lesssim \gamma\kappa\sqrt{\mathrm{sr}(X)} + \frac{1}{d_w^{C_2}} + \sigma\sqrt{s}. \tag{19}$$

Conversely, let $(W_1^{\lambda=0}(t), ..., W_L^{\lambda=0}(t))$ be the weight matrices of a deep linear network trained in the setting of Theorem 2.3 with $\lambda = 0$. Then, for any $\beta > 0$, there exists an iteration T such that the reconstruction error $\|W_{L:1}^{\lambda=0}(t)Y - X\|_{\mathsf{F}} \leq \beta\|X\|_{\mathsf{F}}$ for all $t > T$. Moreover, with high probability, the test error satisfies

$$\|W_{L:1}^{\lambda=0}(t)y - x\| \gtrsim \sigma\left(\sqrt{\frac{d(m-s)}{m}} - \sqrt{s}\right) - \beta\kappa\sqrt{\mathrm{sr}(X)}\|y\|. \tag{20}$$

The benefit of weight decay can be deduced from the qualitative behavior of the two bounds: on one hand, the error in Equation (19) can be driven arbitrarily close to $\sigma\sqrt{s}$ – which is unimprovable in general – by choosing γ sufficiently small and d_w sufficiently large. On the other hand, suppose that ${}^{(m-s)}/_{m} = \Omega(1)$ (a standard regime in compressed sensing tasks): in that case, training without weight decay *always* incurs a test error of at least $\sigma\sqrt{d} - \beta\kappa\sqrt{\mathrm{sr}(X)}\|y\|$. For high-dimensional problem instances, this bound is only vacuous if β scales with the misspecification σ, the ambient dimension $\sqrt{d}$, or both – in other words, the lower bound (20) can be significantly larger than (19) unless $W_{L:1}^{\lambda=0}(t)$ is a poor fit to the training data.

3 Numerical experiments

In this section, we present numerical experiments that corroborate our theoretical findings and examine the sensitivity of the learned mapping to different parameters: the dimension of the latent subspace s (Section 3.1, the depth of the neural network L (Section 3.2), and the regularization strength λ (Section 3.3). In our experiments, we track the regression and "off-subspace" errors across t:

$$\frac{\|W_{L:1}(t)Y - X\|_{\mathsf{F}}}{\|X\|_{\mathsf{F}}} \quad \text{and} \quad \|W_{L:1}(t)P^{\perp}_{\mathrm{range}(Y)}\|_{\mathsf{op}}.$$

Experimental setup. For each experiment shown in Figures 1a and 2 to 4, we generate the measurement matrix A by sampling a random Gaussian matrix $G \in \mathbb{R}^{m \times d}$ and setting $A := \frac{1}{\sqrt{m}} G$; such matrices satisfy Assumption 2.1 with high probability as long as $m \gtrsim s \log(d)$ [56]. To form the subspace basis matrix R, we calculate the QR factorization of a $d \times s$ random Gaussian matrix and keep the orthogonal factor. Finally, we generate the signal matrix $X \in \mathbb{R}^{d \times n}$ as $X = RZ$, where $Z \in \mathbb{R}^{s \times n}$ is a full row-rank matrix of subspace coefficients. Given a target condition number κ for X, we generate Z via its SVD: we sample the left and right singular factors at random and arrange its singular values uniformly in the interval $[\frac{1}{\kappa}, 1]$. All our experiments use step sizes that are covered by our theory but do not necessarily correspond to the value suggested by Theorem 2.3 (see Remark 2.5). Similarly, each experiment uses a number of iterations that is sufficiently large but not necessarily equal to T. Finally, all weight decay parameters used correspond to a valid $\gamma \in (0, 1)$, but for the sake of simplicity we specify λ directly.

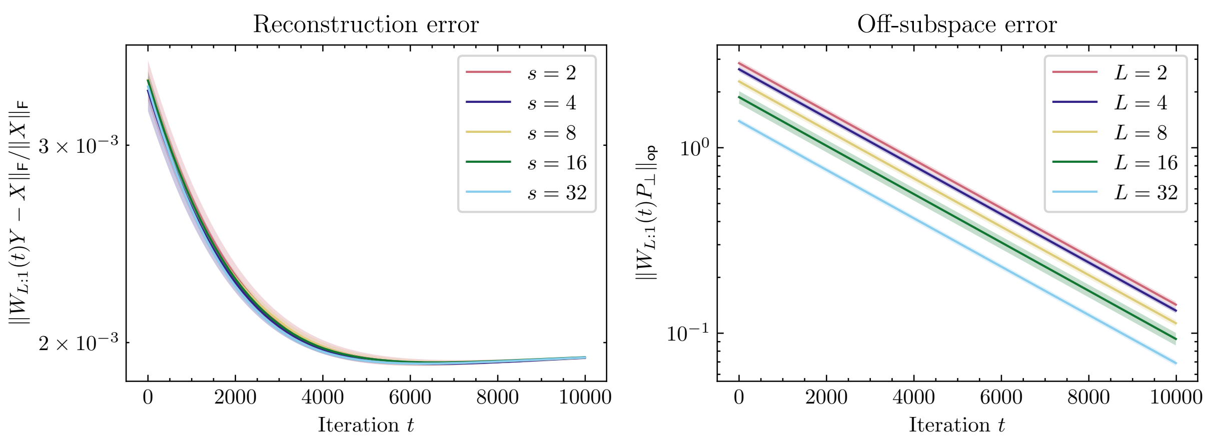


Figure 2: Comparing the training error of a deep linear neural network for data of varying subspace dimensions s using constant stepsize $\eta = 1/10$ and weight decay $\lambda = 10^{-3}$. The lines are the median over 10 runs with independently sampled training data and weight initializations. The shaded region indicates one standard deviation around the median. See Section 3.1 for details.

3.1 Impact of latent subspace dimension s

The statement of Theorem 2.3 suggests that the size of the subspace s does not affect the rate of (on-subspace) convergence or the error achieved after T iterations. To verify this numerically, we generate several synthetic datasets with varying subspace dimension $s \in \{2, 4, 8, 16, 32\}$, $m = 128$, $d = 256$ and perfectly conditioned data (i.e., $\kappa = 1$). For each dataset, we train a deep linear network of width $d_w = 512$ using $\eta = 1/10$ and $\lambda = 10^{-3}$ and compute the median reconstruction and off-subspace errors and standard deviation over 10 independent runs, with each run using $n = 1000$ independently drawn samples. The results, depicted in Figure 2, suggest that the errors decay at the same rate; in the case of the reconstruction error, the differences in magnitude are negligible, while the off-subspace errors differ by a constant offset across subspace dimensions.

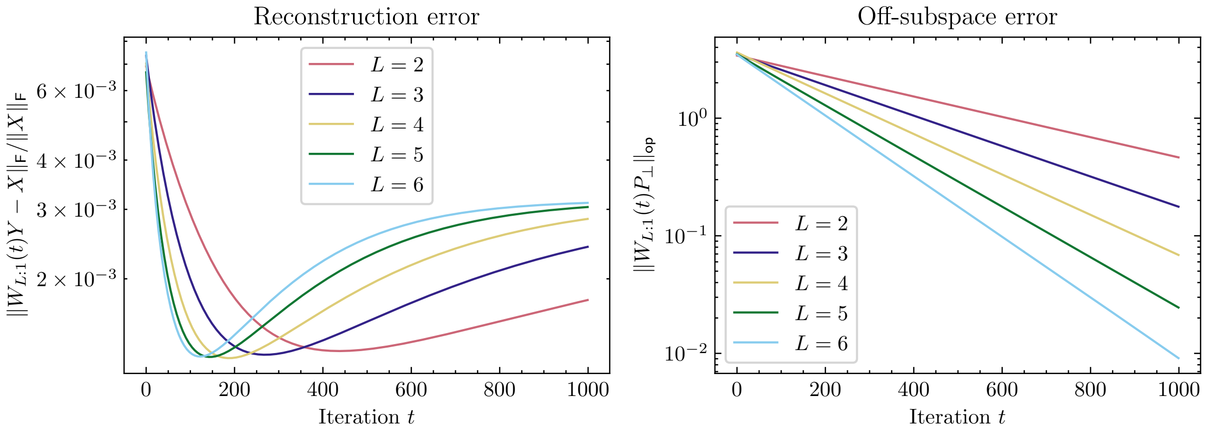


Figure 3: Normalized regression error and off-subspace error for deep linear nets of varying depths L, trained with gradient descent using constant stepsize $\eta = 1/10$ and weight decay parameter $\lambda = 10^{-4}$. While the regression error drops to similar levels for all depths, larger L confers a clear advantage with respect to the off-subspace error. See Section 3.2 for details.

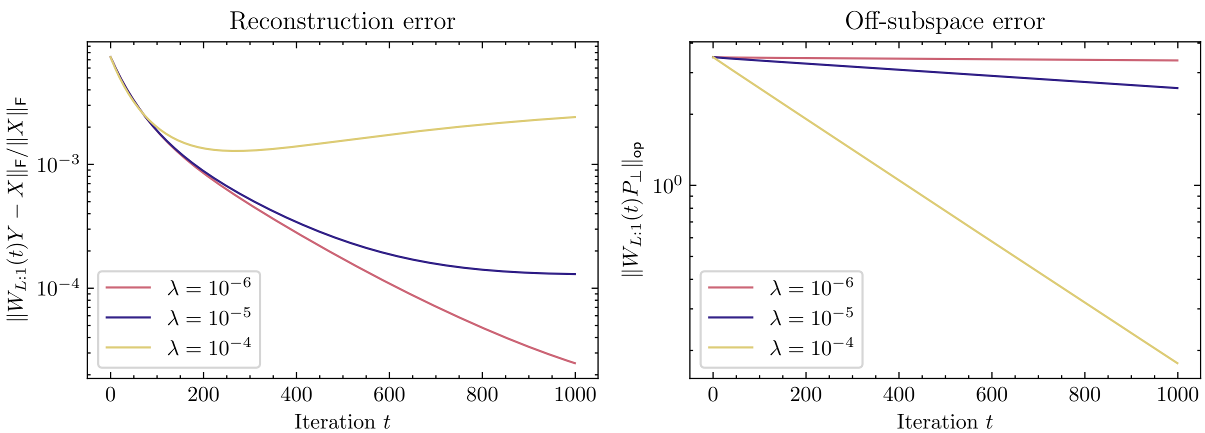


Figure 4: Normalized regression error and off-subspace errors for deep linear nets trained with gradient descent with stepsize $\eta = 1/10$ and varying levels of weight decay λ. While high levels of weight decay reduce the off-subspace error faster, they lead to larger regression error. See Section 3.3 for details.

3.2 Impact of neural network depth L

In our next experiment, we examine how the neural network depth, L, affects convergence and generalization. We generate a dataset with subspace dimension $s = 4$, measurement dimension $m = 32$, signal dimension $d = 64$ and $n = 1000$ samples (using perfectly condition data; i.e., $\kappa = 1$) and train a deep linear network of width $d_w = 1000$ using gradient descent. We use the same stepsize $\eta = 10^{-1}$ and weight decay parameter $\lambda = 10^{-4}$ across all configurations.

The results for both quantities of interest are depicted in Figure 3. The regression error first drops to similar levels, for all depths, before it starts increasing and plateauing at roughly 20λ. However, higher depth L confers a clear advantage with respect to the off-subspace error.

3.3 Impact of weight decay parameter λ

Our next experiment examines the impact of the weight decay parameter λ. We use a similar setup as in Section 3.2, where $s = 4$, $m = 32$ and $d = 64$ with $n = 1000$ samples, and train neural networks of width $d_w = 1000$ and depth $L = 3$; see Figure 4. As Theorem 2.3 suggests, larger weight decay values lead to larger regression errors (approximately $10 \cdot \lambda$) but faster decaying off-subspace errors.

4 Limitations and future directions

Depth and generalization. Our experiments in Figure 3 suggest that depth is beneficial for both the regression and the "off-subspace" errors: larger depth, at least up to a certain point, leads to faster convergence. This phenomenon is not covered by our main theoretical result, but constitutes an interesting direction for future work.

Near-singular matrices and conditioning. Our main result (Theorem 2.3) does not provide meaningful insights for *approximately* low-rank data; e.g., inputs X that can be decomposed as the sum of a well-conditioned low-rank component and a full-rank component with relatively small singular values, a pervasive property in data science applications [58]. For such inputs, it is plausible that gradient descent with weight decay is able to rapidly converge to a solution mapping that provides a good approximation to the "low-rank" component of the input X. We leave such an investigation to future work.

Adaptivity of deep non-linear networks. Our experiments in Figure 1b suggest that weight decay can lead to robust solutions beyond the simple linear inverse problem setting. In particular, a natural next step would be to study the training dynamics of ℓ_2-regularized gradient descent for deep networks with several linear layers and a ReLU layer (as considered in [15]) under the assumption that the input data is generated by the "union-of-subspaces" model used in Figure 1b.

Acknowledgements

SP gratefully acknowledges the support of the NSF Graduate Research Fellowship Program NSF DGE-2140001. VC and RW gratefully acknowledge the support of NSF DMS-2023109, the NSF-Simons National Institute for Theory and Mathematics in Biology (NITMB) through NSF (DMS-2235451) and Simons Foundation (MP-TMPS-00005320), and the Margot and Tom Pritzker Foundation. FK gratefully acknowledges the support of the German Science Foundation (DFG) in the context of the priority program Theoretical Foundations of Deep Learning (project KR 4512/6-1). HL and FK gratefully acknowledge the support of the Munich Center for Machine Learning (MCML).

References

[1] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing (3rd Edition)*. Prentice-Hall, Inc., USA, 2006. ISBN 013168728X.

[2] Michael Lustig, David L Donoho, Juan M Santos, and John M Pauly. Compressed sensing mri. *IEEE signal processing magazine*, 25(2):72–82, 2008.

[3] Michael T McCann, Kyong Hwan Jin, and Michael Unser. Convolutional neural networks for inverse problems in imaging: A review. *IEEE Signal Processing Magazine*, 34(6):85–95, 2017.

[4] Anuroop Sriram, Jure Zbontar, Tullie Murrell, Aaron Defazio, C Lawrence Zitnick, Nafissa Yakubova, Florian Knoll, and Patricia Johnson. End-to-end variational networks for accelerated mri reconstruction. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23*, pages 64–73. Springer, 2020.

[5] Michael Elad, Bahjat Kawar, and Gregory Vaksman. Image denoising: The deep learning revolution and beyond—a survey paper. *SIAM Journal on Imaging Sciences*, 16(3):1594–1654, 2023.

[6] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 417–424, 2000.

[7] Weize Quan, Jiaxi Chen, Yanli Liu, Dong-Ming Yan, and Peter Wonka. Deep learning-based image and video inpainting: A survey. *International Journal of Computer Vision*, 132(7): 2367–2400, 2024.

[8] Jonathan Scarlett, Reinhard Heckel, Miguel RD Rodrigues, Paul Hand, and Yonina C Eldar. Theoretical perspectives on deep learning methods in inverse problems. *IEEE journal on selected areas in information theory*, 3(3):433–453, 2022.

[9] Vegard Antun, Francesco Renna, Clarice Poon, Ben Adcock, and Anders C Hansen. On instabilities of deep learning in image reconstruction and the potential costs of ai. *Proceedings of the National Academy of Sciences*, 117(48):30088–30095, 2020.

[10] Mohammad Zalbagi Darestani, Akshay S Chaudhari, and Reinhard Heckel. Measuring robustness in deep learning based compressive sensing. In *International Conference on Machine Learning*, pages 2433–2444. PMLR, 2021.

[11] Martin Genzel, Jan Macdonald, and Maximilian März. Solving inverse problems with deep neural networks–robustness included? *IEEE transactions on pattern analysis and machine intelligence*, 45(1):1119–1134, 2022.

[12] Anselm Krainovic, Mahdi Soltanolkotabi, and Reinhard Heckel. Learning provably robust estimators for inverse problems via jittering. *Advances in Neural Information Processing Systems*, 36, 2024.

[13] Anders Krogh and John A Hertz. A simple weight decay can improve generalization. In *Proceedings of the 4th International Conference on Neural Information Processing Systems*, pages 950–957, 1991.

[14] Siegfried Bos and E Chug. Using weight decay to optimize the generalization ability of a perceptron. In *Proceedings of International Conference on Neural Networks (ICNN'96)*, volume 1, pages 241–246. IEEE, 1996.

[15] Suzanna Parkinson, Greg Ongie, and Rebecca Willett. Linear neural network layers promote learning single-and multiple-index models. *arXiv preprint arXiv:2305.15598*, 2023.

[16] Peter L Bartlett. For valid generalization, the size of the weights is more important than the size of the network. In *Proceedings of the 9th International Conference on Neural Information Processing Systems*, pages 134–140, 1996.

[17] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

[18] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Conference on Learning Theory*, pages 1376–1401. PMLR, 2015.

[19] Colin Wei, Jason D Lee, Qiang Liu, and Tengyu Ma. Regularization matters: generalization and optimization of neural nets vs their induced kernel. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 9712–9724, 2019.

[20] Amit Daniely and Elad Granot. Generalization bounds for neural networks via approximate description length. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 13008–13016, 2019.

[21] Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. *Information and Inference: A Journal of the IMA*, 9(2):473–504, 2020.

[22] Suzanna Parkinson, Greg Ongie, Rebecca Willett, Ohad Shamir, and Nathan Srebro. Depth separation in norm-bounded infinite-width neural networks. *Conference on Learning Theory (COLT), arXiv preprint arXiv:2402.08808*, 2024.

[23] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

[24] Guodong Zhang, Chaoqi Wang, Bowen Xu, and Roger Grosse. Three mechanisms of weight decay regularization. In *International Conference on Learning Representations*, 2019.

[25] Francesco D'Angelo, Maksym Andriushchenko, Aditya Varre, and Nicolas Flammarion. Why do we need weight decay in modern deep learning? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2023.

[26] Pedro Savarese, Itay Evron, Daniel Soudry, and Nathan Srebro. How do infinite width bounded norm networks look in function space? In *Conference on Learning Theory*, pages 2667–2690. PMLR, 2019.

[27] Greg Ongie, Rebecca Willett, Daniel Soudry, and Nathan Srebro. A function space view of bounded norm infinite width relu nets: The multivariate case. In *International Conference on Learning Representations (ICLR 2020)*, 2019.

[28] Chao Ma, Lei Wu, and Weinan E. The barron space and the flow-induced function spaces for neural network models. *Constructive Approximation*, 55(1):369–406, 2022.

[29] Rahul Parhi and Robert D Nowak. Near-minimax optimal estimation with shallow relu neural networks. *IEEE Transactions on Information Theory*, 69(2):1125–1140, 2022.

[30] Etienne Boursier and Nicolas Flammarion. Penalising the biases in norm regularisation enforces sparsity. *Advances in Neural Information Processing Systems*, 36:57795–57824, 2023.

[31] Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017.

[32] Arthur Jacot. Implicit bias of large depth networks: A notion of rank for nonlinear functions. *International Conference on Learning Representations*, 2023.

[33] Seijin Kobayashi, Yassir Akram, and Johannes Von Oswald. Weight decay induces low-rank attention layers. *arXiv preprint arXiv:2410.23819*, 2024.

[34] Fanhua Shang, Yuanyuan Liu, Fanjie Shang, Hongying Liu, Lin Kong, and Licheng Jiao. A unified scalable equivalent formulation for schatten quasi-norms. *Mathematics*, 8(8):1325, August 2020. ISSN 2227-7390. doi: 10.3390/math8081325.

[35] Zhen Dai, Mina Karzand, and Nathan Srebro. Representation costs of linear neural networks: Analysis and design. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 26884–26896. Curran Associates, Inc., 2021.

[36] Simon Du and Wei Hu. Width provably matters in optimization for deep linear neural networks. In *International Conference on Machine Learning*, pages 1655–1664. PMLR, 2019.

[37] Wei Hu, Lechao Xiao, and Jeffrey Pennington. Provable benefit of orthogonal initialization in optimizing deep linear networks. In *International Conference on Learning Representations*, 2020.

[38] Sanjeev Arora, Nadav Golowich, Noah Cohen, and Wei Hu. A convergence analysis of gradient descent for deep linear neural networks. In *International Conference on Learning Representations*, 2019.

[39] Gabin Maxime Nguegnang, Holger Rauhut, and Ulrich Terstiege. Convergence of gradient descent for learning linear neural networks. *Advances in Continuous and Discrete Models*, 2024(1):23, 2024.

[40] Ziqing Xu, Hancheng Min, Salma Tarmoun, Enrique Mallada, and René Vidal. Linear convergence of gradient descent for finite width over-parametrized linear networks with general initialization. In *International Conference on Artificial Intelligence and Statistics*, pages 2262–2284. PMLR, 2023.

[41] Ohad Shamir. Exponential convergence time of gradient descent for one-dimensional deep linear neural networks. In *Conference on Learning Theory*, pages 2691–2713. PMLR, 2019.

[42] Kenji Kawaguchi. Deep learning without poor local minima. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 586–594, 2016.

[43] Thomas Laurent and James von Brecht. Deep linear networks with arbitrary loss: All local minima are global. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2902–2907. PMLR, 10–15 Jul 2018.

[44] Gauthier Gidel, Francis Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in linear neural networks. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 3202–3211, 2019.

[45] Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. In *International Conference on Learning Representations*, 2019.

[46] Armin Eftekhari. Training linear neural networks: Non-local convergence and complexity results. In *International Conference on Machine Learning*, pages 2836–2847. PMLR, 2020.

[47] Bubacarr Bah, Holger Rauhut, Ulrich Terstiege, and Michael Westdickenberg. Learning deep linear neural networks: Riemannian gradient flows and convergence to global minimizers. *Information and Inference: A Journal of the IMA*, 11(1):307–353, 02 2021. ISSN 2049-8772. doi: 10.1093/imaiai/iaaa039.

[48] Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit bias of sgd for diagonal linear networks: a provable benefit of stochasticity. *Advances in Neural Information Processing Systems*, 34:29218–29230, 2021.

[49] Arthur Jacot, François Ged, Berfin Şimşek, Clément Hongler, and Franck Gabriel. Saddle-to-saddle dynamics in deep linear networks: Small initialization training, symmetry, and sparsity. *arXiv preprint arXiv:2106.15933*, 2021.

[50] Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *International Conference on Machine Learning*, pages 244–253. PMLR, 2018.

[51] Aitor Lewkowycz and Guy Gur-Ari. On the training dynamics of deep networks with ℓ_2 regularization. *Advances in Neural Information Processing Systems*, 33:4790–4799, 2020.

[52] Can Yaras, Peng Wang, Wei Hu, Zhihui Zhu, Laura Balzano, and Qing Qu. The law of parsimony in gradient descent for learning deep linear networks. *arXiv preprint arXiv:2306.01154*, 2023.

[53] Can Yaras, Peng Wang, Laura Balzano, and Qing Qu. Compressible dynamics in deep overparameterized low-rank learning & adaptation. *arXiv preprint arXiv:2406.04112*, 2024.

[54] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[55] Zihan Wang and Arthur Jacot. Implicit bias of SGD in L_2-regularized linear DNNs: One-way jumps from high to low rank. In *The Twelfth International Conference on Learning Representations*, 2024.

[56] Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. Birkhäuser Basel, 2013.

[57] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034, 2015. doi: 10.1109/ICCV.2015.123.

[58] Madeleine Udell and Alex Townsend. Why are big data matrices approximately low rank? *SIAM Journal on Mathematics of Data Science*, 1(1):144–160, 2019. doi: 10.1137/18M1183480.

[59] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[60] Davis Gilton, Greg Ongie, and Rebecca Willett. Neumann networks for linear inverse problems in imaging. *IEEE Transactions on Computational Imaging*, 6:328–343, 2020. doi: 10.1109/TCI.2019.2948732.

A Main result and proof

This section presents the formal version of the main result and the full proof. We start with fixing some notation and assumptions in Appendix A.1, and then state the main result in Appendix A.2. Appendix A.3 shows some general lemmas used in multiple proof steps. We show in Appendix A.4 that certain properties hold at initialization. The main proof then involves three steps. First, we prove by induction in Appendix A.5 that the regression error rapidly decreases during an initial phase of gradient descent. Second, in Appendix A.6, we again use induction to show that the regression error remains small during a subsequent phase of gradient descent. Third, in Appendix A.7 we show that the "off-subspace" error becomes small during this period. We conclude by showing that this method is robust at test time in Appendix A.8.

Let us note that Appendices A.3 to A.5 are based on the proof in [36] of the convergence of gradient descent for the convex problem $\min_W \|WY - X\|_{\mathsf{F}}$. Because of the additional regularization term in our setting, the proof is significantly different. For example, we cannot prove that the error converges towards 0 or stays small for all iterations. Instead, we show in Appendices A.6 and A.7 that the error stays less than $O(\lambda)$ for many iterations, during which time the "off-subspace" error shrinks, leading to good generalization.

A.1 Preliminaries

A.1.1 Notation

We first establish some notation; let

$$W_{j:i} = \prod_{t=i}^{j} W_t = \begin{cases} W_j W_{j-1} \dots W_i, & \text{if } i \le j, \\ I, & \text{otherwise}; \end{cases} \tag{21a}$$

$$U = d_w^{-\frac{L-1}{2}} m^{-\frac{1}{2}} W_{L:1} Y. \tag{21b}$$

$$\Phi = U - X. \tag{21c}$$

The matrix U corresponds to the network predictions, while Φ corresponds to the matrix of training residuals. To refer to a matrix at iteration t of gradient descent, we write $W_i(t)$, $W_{j:i}(t), U(t), \Phi(t)$, etc. When convenient, we write

$$d_i = \begin{cases} d_w & \text{for } 2 \le i \le L \\ m & \text{for } i = 1. \end{cases} \tag{22}$$

With this notation at hand, our loss function becomes

$$f(W_1, \dots, W_L) = \frac{1}{2}\|\Phi\|_{\mathsf{F}}^2 + \frac{\lambda}{2}\sum_{j=1}^{L}\left\|\frac{1}{d_i}W_i\right\|_{\mathsf{F}}^2 \tag{23}$$

We shall also write $C_{\mathsf{prod}}^{(i)}$ and C_{prod} for the following products appearing in our proofs:

$$C_{\mathsf{prod}}^{(i)} := \prod_{\substack{j=1 \\ j\neq i}}^{L}\left(1 - \frac{\eta\lambda}{d_i}\right) \quad \text{and} \quad C_{\mathsf{prod}} := \prod_{i=1}^{L}\left(1 - \frac{\eta\lambda}{d_i}\right). \tag{24}$$

Finally, we let $\mathrm{sr}(X) \in [1, \mathrm{rank}(X)]$ denote the *stable rank* of X, defined as

$$\mathrm{sr}(X) := \left(\frac{\|X\|_{\mathsf{F}}}{\|X\|_{\mathsf{op}}}\right)^2 = \sum_{i\geq 1}\left(\frac{\sigma_i(X)}{\sigma_1(X)}\right)^2. \tag{25}$$

A.1.2 Initialization

Assumption A.1 (Initialization)**.** All the weight matrices W_ℓ are initialized according to:

$$(W_\ell)_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$$

with dimensions $W_1 \in \mathbb{R}^{d_w \times m}$, $W_2, \dots, W_{L-1} \in \mathbb{R}^{d_w \times d_w}$, and $W_L \in \mathbb{R}^{d \times d_w}$.

This is the same as Assumption 2.2 since in the optimization problem in Equation (23) we have explicitly pulled out the normalization factor that comes from the "fan-in" initialization.

A.1.3 Gradient Descent Updates

The gradient of the regression error with respect to W_i is equal to

$$\nabla_{W_i}\left[\frac{1}{2}\|\Phi\|_{\mathsf{F}}^2\right] = d_w^{-\frac{L-1}{2}} m^{-\frac{1}{2}} W_{L:i+1}^{\mathsf{T}} \Phi Y^{\mathsf{T}} W_{i-1:1}^{\mathsf{T}}. \tag{26}$$

The gradient of the ℓ_2-regularization term is

$$\nabla_{W_i}\left[\frac{\lambda}{2}\left\|\frac{1}{\sqrt{d_i}} W_i\right\|_{\mathsf{F}}^2\right] = \frac{\lambda}{d_i} W_i. \tag{27}$$

Hence the gradient descent iteration is as follows:

$$W_i(t+1) = \left(1 - \frac{\eta\lambda}{d_i}\right) W_i(t) - \eta d_w^{-\frac{L-1}{2}} m^{-\frac{1}{2}} W_{L:i+1}(t)^{\mathsf{T}} \Phi(t) Y^{\mathsf{T}} W_{i-1:1}(t)^{\mathsf{T}}, \quad \text{for } 1 \leq i \leq L. \tag{28}$$

A.1.4 Simplifying the number of samples

We may assume that we have exactly s input samples for the purpose of analysis. Indeed, Claim 1 below shows that the gradient descent trajectories remain unchanged when the number of samples $n > s$.

Claim 1. Without loss of generality, we may assume $Z \in \mathbb{R}^{s \times s}$, with $\mathrm{rank}(Z) = s$.

Proof. Since $X = RZ$ where $Z \in \mathbb{R}^{s \times n}$ and $n \geq s$, the economic SVD of Z yields

$$X = RU_Z \Sigma_Z V_Z^{\mathsf{T}}, \quad U_Z \in O(s),\ \Sigma_Z = \mathbf{diag}(\sigma_1, \dots, \sigma_s),\ V_Z \in O(n,s).$$

Since the Frobenius norm is unitarily invariant,

$$\begin{aligned}
\|\Phi\|_{\mathsf{F}} &= \|d_w^{-\frac{L-1}{2}} m^{-\frac{1}{2}} W_{L:1} Y - X\|_{\mathsf{F}} \\
&= \|(d_w^{-\frac{L-1}{2}} m^{-\frac{1}{2}} W_{L:1} AR - R) U_Z \Sigma_Z V_Z^{\mathsf{T}}\|_{\mathsf{F}} \\
&= \|d_w^{-\frac{L-1}{2}} m^{-\frac{1}{2}} W_{L:1} ARU_Z \Sigma_Z - RU_Z \Sigma_Z\|_{\mathsf{F}}.
\end{aligned}$$

Moreover,

$$
\begin{aligned}
\nabla_{W_i}\left[\frac{1}{2}\|\Phi\|_{\mathsf{F}}^2\right] &= d_w^{-\frac{L-1}{2}} m^{-\frac{1}{2}} W_{L:i+1}^{\mathsf{T}} \Phi Y^{\mathsf{T}} W_{i-1:1}^{\mathsf{T}} \\
&= d_w^{-\frac{L-1}{2}} m^{-\frac{1}{2}} W_{L:i+1}^{\mathsf{T}} (d_w^{-\frac{L-1}{2}} m^{-\frac{1}{2}} W_{L:1} A - I) R U_Z \Sigma_Z \underbrace{V_Z^{\mathsf{T}} V_Z}_{I_s} \Sigma_Z U_Z^{\mathsf{T}} R^{\mathsf{T}} A^{\mathsf{T}} W_{i-1:1}^{\mathsf{T}} \\
&= d_w^{-\frac{L-1}{2}} m^{-\frac{1}{2}} W_{L:i+1}^{\mathsf{T}} (d_w^{-\frac{L-1}{2}} m^{-\frac{1}{2}} W_{L:1} A - I) R U_Z \Sigma_Z^2 U_Z^{\mathsf{T}} R^{\mathsf{T}} A W_{i-1:1}^{\mathsf{T}}.
\end{aligned}
$$

Thus, without loss of generality, we can assume that $X = RU_Z\Sigma_Z \in \mathbb{R}^{d\times s}$ since this assumption does not change the gradient descent trajectory or value of the loss function. ☐

Throughout the remainder of this section, we will assume that $n = s$.

A.2 Main result

Given the above assumptions and notation, we can now state the formal version of our main result.

Theorem A.2. *Let Assumptions 2.1 and 2.2 hold with $\delta = \frac{1}{10}$. Furthermore, suppose the following conditions are true:*

$$
\lambda \le \frac{L\sigma_{\min}^2(X)}{400\cdot 35}, \quad d_w \gtrsim d\cdot \mathrm{sr}(X)\cdot \mathrm{poly}(L,\kappa), \quad \eta \le \frac{m}{L\sigma_{\max}^2(X)}, \quad \text{and} \quad \lambda = \gamma\sigma_{\min}^2(X)\sqrt{\frac{m}{d}}, \tag{29}
$$

where $\gamma \in (0,1]$ is a user-specified accuracy parameter. Moreover, define the times

$$
\tau = \inf\left\{ t\in\mathbb{N} \mid \|\Phi(t)\|_{\mathsf{F}} \le \frac{80\gamma\|X\|_{\mathsf{F}}}{L}\right\}, \tag{30a}
$$

$$
T = \frac{2\log(d_w)\sqrt{dm}}{\eta\gamma\sigma_{\min}^2(X)}. \tag{30b}
$$

Then with probability of at least $1 - c_1 e^{-c_2 d}$ over the random initialization,

$$
\|W_{L:1}(t+1)Y - X\|_{\mathsf{F}} \le \begin{cases} \left(1 - \frac{\eta L\sigma_{\min}^2(X)}{32m}\right)\|W_{L:1}(t)Y - X\|_{\mathsf{F}}, & t<\tau; \\ C_1\gamma\|X\|_{\mathsf{F}}, & \tau\le t\le T. \end{cases} \tag{31}
$$

$$
\|W_{L:1}(T)P_{\mathrm{range}(Y)}^{\perp}\|_{\mathsf{op}} \le \left(\frac{1}{d_w}\right)^{C_2}, \tag{32}
$$

where c_1, c_2, C_1 and C_2 are positive universal constants.

Remark A.3. Throughout Appendices A.3 to A.7, Assumptions 2.1 and 2.2 and Equation (29) are in force.

Remark A.4. The condition $\lambda \le L\sigma_{\min}^2(X)/400\cdot 35$ is automatically satisfied for small enough γ.

A.3 Lemmas used for the proof

The following lemma bounds the deviation of $W_i(t)$ from $\left(1 - \frac{\eta\lambda}{d_i}\right)^t W_i(0)$.

Lemma A.5. *For any $i \in [L]$, any $t \in \mathbb{N}$, and any matrix norm $\|\cdot\|$, we have*

$$\begin{aligned}
&\|W_i(t) - \left(1 - \frac{\eta\lambda}{d_i}\right)^t W_i(0)\| \\
&\leq \eta d_w^{-\frac{L-1}{2}} m^{-\frac{1}{2}} \sum_{j=0}^{t-1} \left(1 - \frac{\eta\lambda}{d_i}\right)^{t-1-j} \|W_{L:i+1}(j)^\mathsf{T} \Phi(j) (W_{i-1:1}(j) Y)^\mathsf{T}\|
\end{aligned}$$

Proof. The proof follows from the update formula for W_i in Equation (28). Writing

$$B_t := d_w^{-\frac{L-1}{2}} m^{-\frac{1}{2}} W_{L:i+1}(t)^\mathsf{T} \Phi(t) Y^\mathsf{T} W_{i-1:1}(t)^\mathsf{T}, \tag{33}$$

we rewrite Equation (28) as the recursion

$$\begin{aligned}
W_i(t) &= \left(1 - \frac{\eta\lambda}{d_i}\right) W_i(t-1) - \eta B_{t-1} \\
&= \left(1 - \frac{\eta\lambda}{d_i}\right)^2 W_i(t-2) - \eta \left(1 - \frac{\eta\lambda}{d_i}\right) B_{t-2} - \eta B_{t-1} \\
&\vdots \\
&= \left(1 - \frac{\eta\lambda}{d_i}\right)^t W_i(0) - \eta \sum_{j=0}^{t-1} \left(1 - \frac{\eta\lambda}{d_i}\right)^{t-1-j} B_j.
\end{aligned}$$

Rearranging, taking norms and applying the triangle inequality yields the result. □

A.3.1 Evolution of Product Matrix

Lemma A.6. *For an arbitrary iteration index t, it holds that*

$$\begin{aligned}
W_{L:1}(t+1) &= C_{\mathsf{prod}} W_{L:1}(t) + E_0(t) \\
&\quad - \eta d_w^{-\frac{L-1}{2}} m^{-\frac{1}{2}} \sum_{i=1}^{L} C_{\mathsf{prod}}^{(i)} W_{L:i+1}(t) W_{L:i+1}^\mathsf{T}(t) \Phi(t) Y^\mathsf{T} W_{i-1:1}^\mathsf{T}(t) W_{i-1:1}(t),
\end{aligned} \tag{34}$$

with $E_0(t)$ containing all $O(\eta^2)$ terms.

Proof. This is essentially the decomposition in [36, Section 5], modified since

$$W_i(t+1) = W_i(t) \left(1 - \frac{\eta\lambda}{d_i}\right) - \eta d_w^{-\frac{L-1}{2}} m^{-\frac{1}{2}} W_{L:i+1}(t)^\mathsf{T} \Phi(t) Y^\mathsf{T} W_{i-1:1}(t)^\mathsf{T}. \tag{35}$$

For the sake of brevity, we do not repeat the argument here. □

Evolution of the Network Outputs and Residuals. Armed with Lemma A.6, we right-multiply both sides of (34) by $d_w^{-\frac{L-1}{2}} m^{-\frac{1}{2}} Y$ to obtain

$$\begin{aligned} U(t+1) &= C_{\mathsf{prod}} U(t) + E(t) \\ &\quad - \eta d_w^{-(L-1)} m^{-1} \sum_{i=1}^{L} C_{\mathsf{prod}}^{(i)} W_{L:i+1}(t) W_{L:i+1}(t)^{\mathsf{T}} \Phi(t) Y^{\mathsf{T}} W_{i-1:1}^{\mathsf{T}}(t) W_{i-1:1}(t) Y \end{aligned} \tag{36}$$

where $E(t) := d_w^{-\frac{L-1}{2}} m^{-\frac{1}{2}} E_0(t) Y$.

Vectorizing both sides and using the identity $\mathbf{vec}(AXB) = (B^{\mathsf{T}} \otimes A)\, \mathbf{vec}(X)$ yields

$$\mathbf{vec}(U(t+1)) = C_{\mathsf{prod}}\, \mathbf{vec}(U(t)) - \eta P(t)\, \mathbf{vec}(\Phi(t)) + \mathbf{vec}(E(t)), \tag{37}$$

where we write $P(t)$ for the following matrix (dropping the time index t for brevity):

$$P = d_w^{-(L-1)} m^{-1} \sum_{i=1}^{L} C_{\mathsf{prod}}^{(i)} \left(Y^{\mathsf{T}} W_{i-1:1}^{\mathsf{T}} W_{i-1:1} Y \right) \otimes \left(W_{L:i+1} W_{L:i+1}^{\mathsf{T}} \right) \in \mathbb{R}^{sd \times sd} \tag{38}$$

We subtract $\mathbf{vec}(X)$ from both sides of Equation (37); using the notation from Equation (38), the result is equal to

$$\begin{aligned} \mathbf{vec}(\Phi(t+1)) &= C_{\mathsf{prod}}\, \mathbf{vec}(U(t)) - \mathbf{vec}(X) - \eta P(t)\, \mathbf{vec}(\Phi(t)) + \mathbf{vec}(E(t)) \\ &= (C_{\mathsf{prod}} - 1)\, \mathbf{vec}(U(t)) + (I - \eta P(t))\, \mathbf{vec}(\Phi(t)) + \mathbf{vec}(E(t)) \end{aligned} \tag{39}$$

Taking the Frobenius norm on both sides of (39) and using the triangle inequality yields

$$\begin{aligned} \|\Phi(t+1)\|_{\mathsf{F}} &\leq \|I - \eta P(t)\|_{\mathsf{op}} \|\Phi(t)\|_{\mathsf{F}} + |C_{\mathsf{prod}} - 1|\, \|U(t)\|_{\mathsf{F}} + \|E(t)\|_{\mathsf{F}} \\ &\leq (1 - \eta \lambda_{\min}(P(t)))\, \|\Phi(t)\|_{\mathsf{F}} + \|U(t)\|_{\mathsf{F}} \eta \lambda \left[\frac{L-1}{d_w} + \frac{1}{m} \right] + \|E(t)\|_{\mathsf{F}}, \end{aligned} \tag{40}$$

as long as $\eta \leq \frac{1}{\lambda_{\max}(P(t))}$, using Lemma B.4 in the second inequality. Intuitively, Equation (40) suggests that bounding the spectrum of P will allow us to get a recursive bound on the norm of the residual.

A.3.2 Bounds on the spectrum of P

In this paragraph, we furnish bounds on the spectrum of P in terms of the spectrum of $W_{L:i+1}$ and $W_{i-1:1} Y$, for $i = 1 \dots L$. In the following lemma, we drop the time index t for simplicity.

Lemma A.7. *We have the following inequalities:*

$$\lambda_{\max}(P) \leq d_w^{-(L-1)} m^{-1} \sum_{i=1}^{L} C_{\mathsf{prod}}^{(i)} \sigma_{\max}^2(W_{i-1:1} Y) \sigma_{\max}^2(W_{L:i+1}); \tag{41}$$

$$\lambda_{\min}(P) \geq d_w^{-(L-1)} m^{-1} \sum_{i=1}^{L} C_{\mathsf{prod}}^{(i)} \sigma_{\min}^2(W_{i-1:1} Y) \sigma_{\min}^2(W_{L:i+1}). \tag{42}$$

Proof. The inequalities are straightforward to prove using the definition of P in Equation (38) and the following facts:

1. The largest (or smallest) eigenvalue of a sum of matrices is bounded above (or below) by the sum of the largest (or smallest) eigenvalues.
2. The eigenvalues of a Kronecker product are the products of the eigenvalues of the individual factors.
3. For any matrix A, $\lambda_{\max}(A^\mathsf{T} A) = \sigma_{\max}^2(A)$.

Using these facts, the result is immediate. $\square$

A.4 Properties at initialization

Let us bound the extremal singular values of $W_{j:1}Y$, $W_{L:i}$, $W_{i:j}$ at initialization and bound $\|\Phi\|_\mathsf{F}$ and $\|U\|_\mathsf{F}$ at initialization.

Lemma A.8. *There are universal constants $c_1, c_2 > 0$ such that*

$$\mathbb{P}\left\{\max_{1\leq i<L} d_w^{-\frac{i}{2}}\sigma_{\max}(W_{i:1}(0)Y) \leq \frac{6}{5}\sigma_{\max}(X)\right\} \geq 1 - c_1 \exp\left(-\frac{c_2 d_w}{L}\right),$$
$$\mathbb{P}\left\{\min_{1\leq i<L} d_w^{-\frac{i}{2}}\sigma_{\min}(W_{i:1}(0)Y) \geq \frac{4}{5}\sigma_{\min}(X)\right\} \geq 1 - c_1 \exp\left(-\frac{c_2 d_w}{L}\right).$$

Proof. Let $U\Sigma V^\mathsf{T}$ be the economic SVD of $Y = AX$; since $X \in \mathrm{range}(R)$, where $\dim(\mathrm{range}(R)) = s$, this implies $U \in O(m, s)$, $\Sigma = \mathbf{diag}(\sigma_1, \dots, \sigma_s)$ and $V \in O(n, s)$. Consequently, for all $1 \leq i < L$ we have

$$\begin{aligned}\sigma_{\max}(W_{i:1}(0)Y) &= \sigma_{\max}(W_{i:1}(0)AX) \\ &\leq \sigma_{\max}(W_{i:1}(0)U) \cdot \sigma_{\max}(\Sigma V^\mathsf{T}) \\ &= \sigma_{\max}(W_{i:1}(0)U)\|U\Sigma V^\mathsf{T}\|_{\mathsf{op}} \\ &\leq \sqrt{1+\delta} \cdot \sigma_{\max}(W_{i:1}(0)U) \cdot \sigma_{\max}(X),\end{aligned} \tag{44}$$

where the last inequality follows from Assumption 2.1. Similarly, we have

$$\begin{aligned}\sigma_{\min}(W_{i:1}(0)Y) &= \sigma_{\min}(W_{i:1}(0)AX) \\ &\geq \sigma_{\min}(W_{i:1}(0)U) \cdot \sigma_{\min}(\Sigma V^\mathsf{T}) \\ &= \sigma_{\min}(W_{i:1}(0)U) \cdot \sigma_{\min}(AX) \\ &\geq \sqrt{1-\delta} \cdot \sigma_{\min}(W_{i:1}(0)U) \cdot \sigma_{\min}(X).\end{aligned} \tag{45}$$

We proceed with bounding the singular values of $W_{i:1}(0)U$. Note that

$$W_1(0)U = \begin{bmatrix} \langle (W_1(0))_{1,:}, U_{:,1}\rangle & \dots & \langle (W_1(0))_{1,:}, U_{:,s}\rangle \\ \langle (W_1(0))_{2,:}, U_{:,1}\rangle & \dots & \langle (W_1(0))_{2,:}, U_{:,s}\rangle \\ \vdots & & \vdots \\ \langle (W_1(0))_{d_w,:}, U_{:,1}\rangle & \dots & \langle (W_1(0))_{d_w,:}, U_{:,s}\rangle \end{bmatrix} \overset{(d)}{=} G \in \mathbb{R}^{d_w \times s}, \quad G_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \tag{46}$$

since any two components are Gaussian and uncorrelated. Indeed, we have that

$$\mathbb{E}\left[\langle (W_1(0))_{i,:}, U_{:,j}\rangle \langle (W_1(0))_{k,:}, U_{:,\ell}\rangle\right] = \mathsf{tr}\Big(U_{:,j}^\mathsf{T}\mathbb{E}\left[(W_1(0))_{i,:}(W_1(0))_{k,:}^\mathsf{T}\right] U_{:,\ell}\Big)$$

$$= \begin{cases} 0, & i \neq k \\ \langle U_{:,j}, U_{:,\ell} \rangle = 0, & i = k \end{cases},$$

using the fact that $W_1(0)$ has isotropic and U has orthogonal columns. We now apply Lemma B.2 with

$$A_1 = W_1(0)U, A_2 = W_2, \dots, A_i = W_i, \;\text{ and }\; n_1 = n_2 = \dots = n_i = d_w.$$

For these parameter choices, we have $\sum_{j=1}^{i} \frac{1}{n_j} = \frac{i}{d_w}$. Thus, for any fixed $y \in \mathbb{S}^{s-1}$ and $i < L$, Lemma B.2 yields

$$\mathbb{P}\left\{ \left| \|W_{i:1}(0)Uy\|^2 - d_w^i \right| \geq \frac{1}{10} d_w^i \right\} \leq c_1 \exp\left(-\frac{c_2 d_w}{i} \right). \tag{47}$$

Taking an ε-net $\mathcal{N}_\varepsilon$ of $\mathbb{S}^{s-1}$ and using [59, Exercise 4.3.4], we obtain for $i < L$

$$\begin{aligned} \sup_{y \in \mathbb{S}^{s-1}} \left| \|W_{i:1}(0)Uy\|^2 - d_w^i \right| &\leq \frac{1}{1-2\varepsilon} \sup_{y \in \mathcal{N}_\varepsilon} \left| \|W_{i:1}(0)Uy\|^2 - d_w^i \right| \\ &\leq d_w^i \cdot \frac{1}{10(1-2\varepsilon)}, \end{aligned} \tag{48}$$

where the last inequality holds with probability at least $1 - c_1 |\mathcal{N}_\varepsilon| \exp\left(-\frac{c_2 d_w}{i}\right)$ as a result of a union bound over $\mathcal{N}_\varepsilon$. Hence for $i < L$,

$$\begin{aligned} \sigma_{\max}^2(W_{i:1}(0)U) &= \sup_{x \in \mathbb{S}^{s-1}} \|W_{i:1}(0)Ux\|^2 \\ &\leq d_w^i + \sup_{x \in \mathbb{S}^{s-1}} \left| \|W_{i:1}(0)Ux\|^2 - d_w^i \right| \\ &\leq d_w^i \left(1 + \frac{1}{10(1-2\varepsilon)} \right). \end{aligned} \tag{49}$$

Similarly for $i < L$,

$$\begin{aligned} \sigma_{\min}^2(W_{i:1}(0)U) &= \inf_{x \in \mathbb{S}^{s-1}} \|W_{i:1}(0)Ux\|^2 \\ &\geq d_w^i - \sup_{x \in \mathbb{S}^{s-1}} \left| \|W_{i:1}(0)Ux\|^2 - d_w^i \right| \\ &\geq d_w^i \left(1 - \frac{1}{10(1-2\varepsilon)} \right). \end{aligned} \tag{50}$$

Letting $\varepsilon = 1/10$ in Equations (49) and (50) and applying the bounds of Equations (44) and (45) shows that the bound holds for each individual i with probability at least

$$\begin{aligned} 1 - c_1 \exp\left\{ -\frac{c_2 d_w}{i} + \log|\mathcal{N}_\varepsilon| \right\} &\geq 1 - c_1 \exp\left\{ -\frac{c_2 d_w}{i} + s \log\left(1 + \frac{2}{\varepsilon}\right) \right\} \\ &\geq 1 - c_1 \exp\left(-\frac{c_2 d_w}{2i} \right), \end{aligned}$$

as long as $d_w \gtrsim Ls$, using the bound [59, Corollary 4.2.13]:

$$|\mathcal{N}_\varepsilon| \leq \left(1 + \frac{2}{\varepsilon}\right)^s.$$

Taking an additional union bound over $i = 1, \dots, L-1$ combined with the condition $d_w \gtrsim Ls \log(L)$ yields the claim. $\square$

Lemma A.9. *There exist constants $c, C > 0$ such that*

$$\mathbb{P}\left\{\max_{1<k\leq j<L} d_w^{-\frac{j-k+1}{2}}\|W_{j:k}(0)\|_{\mathsf{op}} \leq \sqrt{\frac{L}{c}}\right\} \geq 1 - \exp\left(-\frac{Cd_w}{L}\right). \tag{51}$$

Proof. Since $W_i \in \mathbb{R}^{n_i \times n_{i-1}} = \mathbb{R}^{d_w \times d_w}$ for all $1 < i < L$, Lemma B.2 implies

$$\mathbb{P}\left\{0.9d_w^{j-k+1}\|y\|^2 \leq \|W_j \ldots W_k y\|^2 \leq 1.1d_w^{j-k+1}\|y\|^2\right\} \geq 1 - 2\exp\left(-\frac{c_1 d_w}{j-k+1}\right).$$

In the following choose $y \in \mathbb{S}^{d_w-1}$ and a small constant $c_2 < c_1$. We can partition $[d_w]$ into $\frac{L}{c_2}$ sets, each of size $\frac{c_2 d_w}{L}$. Therefore we can write

$$[d_w] = S_1 \cup \cdots \cup S_{\frac{L}{c_2}}.$$

Let $\mathsf{supp}(u) := \{i \mid u_i \neq 0\}$ and $U_{S_\ell} := \{u \in \mathbb{S}^{d_w-1} \mid \mathsf{supp}(u) \subset S_\ell\}$. Taking a $\frac{1}{2}$-net $\mathcal{N}_\ell$ of U_{S_ℓ}, we obtain:

$$\|W_{j:k}(0)u_\ell\| \leq \sqrt{1.1}d_w^{\frac{j-k+1}{2}} \cdot \frac{3}{2} \leq 2d_w^{\frac{j-k+1}{2}}, \text{ for all } u_\ell \in U_{S_\ell},$$

with the probability of failure at most

$$\begin{aligned}|\mathcal{N}_\ell|\exp\left(-\frac{c_1 d_w}{j-k+1}\right) &\leq \log\left(1 + \frac{2}{1/2}\right)^{|S_\ell|}\exp\left(-\frac{c_1 d_w}{j-k+1}\right)\\ &\leq \exp\left(-\frac{c_1 d_w}{j-k+1} + \frac{c_2 d_w}{L}\log(5)\right)\\ &\leq \exp\left(-\frac{d_w}{L}\left(c_1 - c_2\log(5)\right)\right).\end{aligned}$$

The above inequality holds for all ℓ at the same time with probability of at least

$$1 - \frac{L}{c_2}\exp\left(-\frac{d_w}{L}(c_1 - c_2\log(5))\right) \geq 1 - \exp\left(-\frac{Cd_w}{L}\right),$$

for some small constant $C > 0$, as long as $d_w \gtrsim L\log\frac{L}{c_2}$ and $c_2 \leq \frac{c_1}{2\log(5)}$ (as a result of a union bound).

Finally, note that we can write any unit vector $y \in \mathbb{S}^{d_w-1}$ as

$$y = \sum_\ell \alpha_\ell u_\ell, \;\; u_\ell \in U_{S_\ell}, \;\; \sum_\ell \alpha_\ell^2 = 1.$$

Using the triangle inequality and conditioning on the previous event, we obtain

$$\|W_{j:i}(0)y\| \leq \sum_\ell \|W_{j:i}(0)\alpha_\ell u_\ell\| \leq 2d_w^{\frac{j-k+1}{2}}\sum_\ell|\alpha_\ell| \leq 2d_w^{\frac{j-k+1}{2}}\sqrt{\frac{L}{c_1}\sum_\ell \alpha_\ell^2} \leq d_w^{j-k+1}\sqrt{\frac{L}{c}}, \tag{52}$$

where the last step is using norm equivalence, and relabeling $c := \frac{c_1}{4}$. This completes the proof of the first display in Lemma A.9.

Finally, to prove Equation (51), we apply the union bound over at most $\binom{L}{2} = O(L^2)$ pairs of indices i, j and use the fact that $d_w \gtrsim L\log(\frac{L}{c_1})$. $\square$

Lemma A.10. *There is a universal constant $C > 0$ such that*

$$\mathbb{P}\left\{\max_{1<i\le L} d_w^{-\frac{L-i+1}{2}}\sigma_{\max}(W_{L:i}(0)) \le \frac{6}{5}\right\} \ge 1 - \exp\left(-\frac{Cd_w}{L}\right), \tag{53a}$$

$$\mathbb{P}\left\{\min_{1<i\le L} d_w^{-\frac{L-i+1}{2}}\sigma_{\min}(W_{L:i}(0)) \ge \frac{4}{5}\right\} \ge 1 - \exp\left(-\frac{Cd_w}{L}\right). \tag{53b}$$

Proof. Since $W_i^{\mathsf{T}} \in \mathbb{R}^{d_w \times d_w}$ for $1 < i < L$ and $W_L^{\mathsf{T}} \in \mathbb{R}^{d_w \times d}$, it follows from Lemma B.2 that

$$\mathbb{P}\left\{\left|\|W_{L:i}^{\mathsf{T}}(0)y\|^2 - d_w^{L-i+1}\right| \ge d_w^{L-i+1}\frac{1}{10}\right\} \le 2\exp\left(-\frac{c_1 d_w}{L-i+1}\right)$$

for any $y \in \mathbb{S}^{d-1}$ and some $c_1 > 0$. Taking an ε-net $\mathcal{N}_\varepsilon$ of $\mathbb{S}^{d-1}$ and using [59, Exercise 4.3.4], we have

$$\sup_{y\in\mathbb{S}^{d-1}}\left|\|W_{L:i}^{\mathsf{T}}(0)y\|^2 - d_w^{L-i+1}\right| \le \frac{1}{1-2\varepsilon}\sup_{y\in\mathcal{N}_\varepsilon}\left|\|W_{L:i}^{\mathsf{T}}(0)y\|^2 - d_w^{L-i+1}\right| \le \frac{d_w^{L-i+1}}{10\cdot(1-2\varepsilon)}, \tag{54}$$

where the last inequality holds with probability at least $1 - 2|\mathcal{N}_\varepsilon|\exp\left\{-\frac{cd_w}{L-i+1}\right\}$ as a result of Lemma B.2 and a union bound over $\mathcal{N}_\varepsilon$. In light of Equation (54), we have

$$\begin{aligned}
\sigma_{\max}^2(W_{L:i}^{\mathsf{T}}(0)) &= \sup_{x\in\mathbb{S}^{d-1}}\|W_{L:i}^{\mathsf{T}}(0)x\|^2 \\
&\le d_w^{L-i+1} + \sup_{x\in\mathbb{S}^{d-1}}\left|\|W_{L:i}^{\mathsf{T}}(0)x\|^2 - d_w^{L-i+1}\right| \\
&\le d_w^{L-i+1}\cdot\left(1 + \frac{1}{10(1-2\varepsilon)}\right).
\end{aligned} \tag{55}$$

At the same time, Equation (54) leads to the lower bound

$$\begin{aligned}
\sigma_{\min}^2(W_{L:i}^{\mathsf{T}}(0)) &= \inf_{x\in\mathbb{S}^{d-1}}\|W_{L:i}^{\mathsf{T}}(0)x\|^2 \\
&\ge d_w^{L-i+1} - \sup_{x\in\mathbb{S}^{d-1}}\left|\|W_{L:i}^{\mathsf{T}}(0)x\|^2 - d_w^{L-i+1}\right| \\
&\ge d_w^{L-i+1}\cdot\left(1 - \frac{1}{10(1-2\varepsilon)}\right).
\end{aligned} \tag{56}$$

Letting $\varepsilon = 0.25$ in Equations (55) and (56) we obtain the bound for each individual i with probability of at least

$$\begin{aligned}
1 - 2\exp\left\{-\frac{cd_w}{L-i+1} + \log|\mathcal{N}_\varepsilon|\right\} &\ge 1 - 2\exp\left\{-\frac{cd_w}{L-i+1} + d\log\left(1+\frac{2}{\varepsilon}\right)\right\} \\
&\ge 1 - 2\exp\left(-\frac{cd_w}{2L}\right),
\end{aligned}$$

as long as $d_w \gtrsim Ld$, using the bound from in [59, Corollary 4.2.13]:

$$|\mathcal{N}_\varepsilon| \le \left(1 + \frac{2}{\varepsilon}\right)^d.$$

To prove Equations (53a) and (53b) we apply a union bound over $1 < i \le L$ and require that $d_w \gtrsim Ld\log(L)$. ☐

Lemma A.11. *At initialization, it holds that*

$$\|\Phi(0)\|_{\mathsf{F}} \leq \left(\frac{6}{5}\sqrt{\frac{d}{m}}+1\right)\|X\|_{\mathsf{F}} \leq \left(\frac{11}{5}\sqrt{\frac{d}{m}}\right)\|X\|_{\mathsf{F}} \quad \text{and} \quad \|U(0)\|_{\mathsf{F}} \leq \frac{6}{5}\sqrt{\frac{d}{m}}\|X\|_{\mathsf{F}} \tag{57}$$

with probability at least $1 - c_1 \exp(-c_2 d)$ *as long as* $m \gtrsim s$ *and* $d_w \gtrsim Lm$.

Proof. Let $\bar{U}\bar{\Sigma}\bar{V}^{\mathsf{T}}$ denote the economic SVD of AX, with $\bar{U} \in O(m, s)$. We have

$$\begin{aligned}
\|U(0)\|_{\mathsf{F}} &= \|d_w^{-\frac{L-1}{2}} m^{-\frac{1}{2}} W_{L:1}(0) Y\|_{\mathsf{F}} \\
&= d_w^{-\frac{L-1}{2}} m^{-\frac{1}{2}} \|W_{L:1}(0) A X\|_{\mathsf{F}} \\
&\leq d_w^{-\frac{L-1}{2}} m^{-\frac{1}{2}} \|W_{L:1}(0)\bar{U}\|_{\mathsf{op}} \|\bar{\Sigma}\bar{V}^{\mathsf{T}}\|_{\mathsf{F}} \\
&\leq \sqrt{1+\delta} \cdot d_w^{-\frac{L-1}{2}} m^{-\frac{1}{2}} \|W_{L:1}(0)\bar{U}\|_{\mathsf{op}} \|X\|_{\mathsf{F}},
\end{aligned}$$

where the last inequality follows from Assumption 2.1 and unitary invariance of the norm. Similarly, we have

$$\|\Phi(0)\|_{\mathsf{F}} = \|U(0) - X\|_{\mathsf{F}} \leq \|U(0)\|_{\mathsf{F}} + \|X\|_{\mathsf{F}}.$$

Consequently, it suffices to bound $d_w^{-\frac{L-1}{2}} m^{-\frac{1}{2}} \|W_{L:1}(0)\bar{U}\|_{\mathsf{op}}$. To that end, we invoke Lemma B.2 with

$$A_1 = W_1(0)\bar{U}, A_2 = W_2, \dots, A_L = W_L, \ \text{ with } \ n_1 = n_2 = \cdots = n_L = d_w, \ \text{and } n_{L+1} = d.$$

For these choices, the failure probability will depend on the term

$$\sum_{i=1}^{L} \frac{1}{n_i} = \frac{L-1}{d_w} + \frac{1}{d} \leq \frac{2}{d},$$

under the assumption $d_w \gtrsim L \cdot d$. Indeed, Lemma B.2 yields (for any fixed $y \in \mathbb{R}^s$):

$$\mathbb{P}\left\{ \left| \|W_{L:1}(0) U y\|^2 - d \cdot d_w^{L-1} \|y\|^2 \right| \geq \frac{1}{10} d \cdot d_w^{L-1} \|y\|^2 \right\} \leq c_1 \exp(-c_2 d). \tag{58}$$

Taking an ε-net of $\mathbb{S}^{s-1}$ and proceeding as in the proof of Lemma A.8, we obtain

$$\|W_{L:1}(0) U\|_{\mathsf{op}}^2 \leq d \cdot d_w^{L-1} \left(1 + \frac{1}{10(1-2\varepsilon)}\right)$$

with probability at least $1 - c_1 \exp\left(-c_2 d + s \log(1 + \frac{2}{\varepsilon})\right) \geq 1 - \exp(-c_2 {d}/{2})$, since $d \geq m \gtrsim s$ for Assumption 2.1 to be valid. Finally, letting $\varepsilon = 1/10 = \delta$, we obtain

$$\sqrt{\frac{1+\delta}{d_w^{L-1} m}} \|W_{L:1}(0)\bar{U}\|_{\mathsf{op}} \leq \frac{6}{5}\sqrt{\frac{d}{m}},$$

as expected. This completes the proof. □

Before we proceed with the proof, we note that a simple union bound shows that all the bounds in Lemmas A.8 to A.11 are fulfilled simultaneously with probability at least $1 - c_1 \exp(-c_2 d)$, for appropriate universal constants $c_1, c_2 > 0$.

A.5 Step 1: Rapid early convergence

The first step of our convergence analysis is showing a sufficient decrease in the regression error until time τ as defined in Equation (30a). We will prove the following theorem in this section.

Theorem A.12. *For all $0 \le t \le \tau$, the following events hold with probability of at least $1 - c_1 e^{-c_2 d}$, where $c_1, c_2 > 0$ are universal constants, over the random initialization:*

$$\mathcal{A}(t) := \left\{ \|\Phi(t+1)\|_{\mathsf{F}} \le \left(1 - \frac{\eta L \sigma_{\min}^2(X)}{16m}\right) \|\Phi(t)\|_{\mathsf{F}} + \frac{5\eta\lambda}{2m}\sqrt{\frac{d}{m}}\|X\|_{\mathsf{F}} \right\} \tag{59a}$$

$$\mathcal{B}(t) := \left\{ \begin{array}{lll} \sigma_{\max}(W_{j:i}(t)) & \le & 2\sqrt{\frac{L}{c}} d_w^{\frac{j-i+1}{2}}, \quad 1 < i \le j < L \\ \sigma_{\max}(W_{i:1}(t)Y) & \le & \frac{5}{4} d_w^{\frac{i}{2}} \cdot \sigma_{\max}(X), \quad 1 \le i < L, \\ \sigma_{\max}(W_{L:i}(t)) & \le & \frac{5}{4} d_w^{\frac{L-i+1}{2}}, \quad 1 < i \le L, \\ \sigma_{\min}(W_{i:1}(t)Y) & \ge & \frac{3}{4} d_w^{\frac{i}{2}} \cdot \sigma_{\min}(X), \quad 1 \le i < L, \\ \sigma_{\min}(W_{L:i}(t)) & \ge & \frac{3}{4} d_w^{\frac{L-i+1}{2}}, \quad 1 \le i < L. \end{array} \right\}, \tag{59b}$$

$$\mathcal{C}(t) := \left\{ \|W_i(t) - \left(1 - \frac{\eta\lambda}{d_i}\right)^t W_i(0)\|_{\mathsf{op}} \lesssim R \mid 1 \le i \le L \right\}, \quad \textit{where } R := \frac{\kappa^2 \sqrt{d\,\mathrm{sr}(X)}}{L}. \tag{59c}$$

We will prove the above theorem by induction, starting with $t = 0$ (Lemma A.13). We then proceed by showing that:

- $\{\mathcal{A}(j)\}_{j<t}$ and $\mathcal{B}(t)$ imply $\mathcal{A}(t)$ (Lemmas A.14 and A.17);
- $\{\mathcal{A}(j), \mathcal{B}(j)\}_{j<t}$ imply $\mathcal{C}(t)$ (Lemma A.19);
- $\mathcal{C}(t)$ implies $\mathcal{B}(t)$ (Lemma A.20).

The proof of Theorem A.12 follows by iterating the above implications until the stopping time τ is reached.

Lemma A.13 (Initialization)**.** *The events $\mathcal{A}(0)$, $\mathcal{B}(0)$ and $\mathcal{C}(0)$ hold with probability at least $1 - c_1 e^{-c_2 d}$, where $c_1, c_2 > 0$ are universal constants.*

Proof. The base case $\mathcal{C}(0)$ is trivial. On the other hand, $\mathcal{B}(0)$ follows from Lemmas A.8, A.9 and A.10. Finally, we show in Lemma A.17 that $\mathcal{B}(t)$ implies $\mathcal{A}(t)$ for all t, including $t = 0$. $\square$

Lemma A.14. *Fix $t \le \tau$ and suppose that $\{\mathcal{A}(j)\}_{j \le t-1}$ and $\{\mathcal{B}(j)\}_{j \le t}$ hold. Then*

$$\|E(t)\|_{\mathsf{F}} \le \frac{17\eta L \sigma_{\min}^2(X)}{1024m} \cdot \|\Phi(t)\|_{\mathsf{F}}. \tag{60}$$

Proof. Note that each term in $E(t)$ is the product of 2 or more terms of the form $\nabla_{W_i} \frac{1}{2}\|\Phi\|_{\mathsf{F}}^2$ and $L-2$ or fewer terms of the form $W_i(t)(1 - \eta\lambda/d_i)$. When ℓ of these terms are from the former category, there are $\binom{L}{\ell}$ ways to choose their indices $(s_1, \dots, s_\ell)$. Each such choice induces a term $C_{(s_1,\dots,s_\ell)}$, defined by

$C_{(s_1,\dots,s_\ell)}$

$$:= \eta^\ell \widetilde{W}_{L:(s_\ell+1)} \left(\nabla_{W_{s_\ell}} \frac{1}{2}\|\Phi\|_{\mathsf{F}}^2\right) \widetilde{W}_{(s_\ell-1):(s_{\ell-1}+1)} \left(\nabla_{W_{s_{\ell-1}}} \frac{1}{2}\|\Phi\|_{\mathsf{F}}^2\right) \dots \left(\nabla_{W_{s_1}} \frac{1}{2}\|\Phi\|_{\mathsf{F}}^2\right) \widetilde{W}_{(s_1-1):1}$$

where we define the products $\widetilde{W}_{i:j}$ as $\widetilde{W}_{i:j} = W_{i:j} \prod_{k=i}^{j} \left(1 - \frac{\eta\lambda}{d_k}\right)$. Each factor of the form $\nabla_{W_k} \frac{1}{2}\|\Phi\|_{\mathsf{F}}^2$ satisfies

$$\begin{aligned}
\|\nabla_{W_k} \frac{1}{2}\|\Phi\|_{\mathsf{F}}^2\|_{\mathsf{F}} &\le d_w^{-\frac{L-1}{2}} m^{-\frac{1}{2}} \|W_{L:(k+1)}(t)\|_{\mathsf{op}} \|\Phi(t)\|_{\mathsf{F}} \|W_{(k-1):1}(t)Y\|_{\mathsf{op}} \\
&\le \frac{5}{4} d_w^{-\frac{L-1}{2}} m^{-\frac{1}{2}} \cdot d_w^{\frac{L-k}{2}} \|\Phi(t)\|_{\mathsf{F}} \frac{5}{4} d_w^{\frac{k-1}{2}} \|X\|_{\mathsf{op}} \\
&= \frac{25}{16\sqrt{m}} \|\Phi(t)\|_{\mathsf{F}} \|X\|_{\mathsf{op}}.
\end{aligned} \tag{61}$$

From $\mathcal{B}(t)$, the factors $W_{(s_\ell-1):(s_{\ell-1}+1)}$ satisfy $\|W_{(s_\ell-1):(s_{\ell-1}+1)}\|_{\mathsf{F}} \le 2\sqrt{\frac{L}{c}} d_w^{\frac{s_\ell - s_{\ell-1} - 1}{2}}$. From $\mathcal{B}(t)$ and Assumption 2.1, we also get $\|W_{(s_1-1):1}Y\|_{\mathsf{F}} \le \frac{5}{4} d_w^{\frac{s_1-1}{2}} \sigma_{\max}(X)$. Similarly, we have $\|W_{L:s_\ell+1}\|_{\mathsf{F}} \le 2\sqrt{\frac{L}{c}} d_w^{\frac{L-s_\ell}{2}}$. Consequently, $C_{(s_1,\dots,s_\ell)}Y$ admits the following bound:

$$\begin{aligned}
&\|C_{(s_1,\dots,s_\ell)}Y\|_{\mathsf{F}} \\
&\le \eta^\ell \left(\prod_{k \notin \{s_1,\dots,s_\ell\}} \left(1 - \frac{\eta\lambda}{d_k}\right) \right) \cdot \left(\frac{25}{16\sqrt{m}} \|\Phi(t)\|_{\mathsf{F}} \|X\|_{\mathsf{op}} \right)^\ell \cdot \left[\frac{5}{4} d_w^{\frac{s_1-1}{2}} \|X\|_{\mathsf{op}} \cdot 2\sqrt{\frac{L}{c}} d_w^{\frac{L-s_\ell}{2}} \prod_{k=1}^{\ell-1} 2\sqrt{\frac{L}{c}} d_w^{\frac{s_{k+1}-s_k-1}{2}} \right].
\end{aligned}$$

Note that the last term equals

$$\frac{5}{4} d_w^{\frac{s_1-1}{2}} \|X\|_{\mathsf{op}} \cdot 2\sqrt{\frac{L}{c}} d_w^{\frac{L-s_\ell}{2}} \prod_{k=1}^{\ell-1} 2\sqrt{\frac{L}{c}} d_w^{\frac{s_{k+1}-s_k-1}{2}} = \frac{5}{4}\|X\|_{\mathsf{op}} \left(2\sqrt{\frac{L}{c}}\right)^\ell \cdot d_w^{\frac{L-\ell}{2}}, \tag{62}$$

and the first term, comprising products for indices different from $\{s_1, \dots, s_\ell\}$, satisfies:

$$\prod_{k \notin \{s_1,\dots,s_\ell\}} \left(1 - \frac{\eta\lambda}{d_k}\right) \le \left(1 - \frac{\eta\lambda}{d_w}\right)^{L-\ell}, \tag{63}$$

since $d_w \ge m$ by assumption. Putting Equations (61) to (63) together, we obtain

$$\begin{aligned}
\|E(t)\|_{\mathsf{F}} &= \|d_w^{-\frac{L-1}{2}} m^{-\frac{1}{2}} E_0(t) Y\|_{\mathsf{F}} \\
&\le \frac{5}{4}\|X\|_{\mathsf{op}} d_w^{-\frac{L-1}{2}} m^{-\frac{1}{2}} \sum_{\ell=2}^{L} \eta^\ell \binom{L}{\ell} \left(1 - \frac{\eta\lambda}{d_w}\right)^{L-\ell} \left(2\sqrt{\frac{L}{c}}\right)^\ell d_w^{\frac{L-\ell}{2}} \left(\frac{25}{16\sqrt{m}} \|\Phi(t)\|_{\mathsf{F}} \|X\|_{\mathsf{op}}\right)^\ell \\
&\le \frac{5}{4}\|X\|_{\mathsf{op}} \sqrt{\frac{d_w}{m}} \sum_{\ell=2}^{L} \left(\frac{C\eta L^{\frac{3}{2}} \|X\|_{\mathsf{op}} \|\Phi(t)\|_{\mathsf{F}}}{\sqrt{m \cdot d_w} \cdot (1 - \frac{\eta\lambda}{d_w})} \right)^\ell \\
&\le \frac{5C\eta L^{\frac{3}{2}} \|X\|_{\mathsf{op}}^2 \|\Phi(t)\|_{\mathsf{F}}}{4m \cdot (1 - \frac{\eta\lambda}{d_w})} \sum_{\ell=1}^{L-1} \left(\frac{C\eta L^{\frac{3}{2}} \|X\|_{\mathsf{op}} \|\Phi(t)\|_{\mathsf{F}}}{(m d_w)^{1/2} (1 - \eta\lambda/d_w)} \right)^\ell,
\end{aligned}$$

where the second to last inequality was obtained from the following bounds:

- for any $j \in \mathbb{N}$, we have $\binom{L}{j} \leq L^j$;
- for any $j \in \mathbb{N}$, we have $(1 - \eta\lambda/d_w)^j \leq 1$;
- finally, we relabel $C := 2\sqrt{\frac{1}{c}} \cdot \frac{25}{16}$ for simplicity.

Note that $\eta\lambda \leq \frac{d_w}{2}$ implies $\eta/(1-\frac{\eta\lambda}{d_w}) \leq 2\eta$. Consequently,

$$\begin{aligned}\frac{2C\eta L^{3/2}\|X\|_{\mathsf{op}}\|\Phi(t)\|_{\mathsf{F}}}{\sqrt{md_w}} &\leq \frac{2CL^{1/2}\sqrt{m}\|\Phi(t)\|_{\mathsf{F}}}{\sqrt{d_w}\sigma_{\max}(X)} \\ &\lesssim \frac{\sqrt{L}\|X\|_{\mathsf{F}}\sqrt{d}}{\sigma_{\max}(X)\sqrt{d_w}} \\ &\lesssim \sqrt{\frac{Ld\,\mathrm{sr}(X)}{d_w}} \\ &\leq \frac{1}{2},\end{aligned}$$

where the first inequality follows from the upper bound on η, the second inequality follows from $\mathcal{A}(0), \ldots, \mathcal{A}(t-1)$, which together with the definition of τ imply that $\|\Phi(t)\|_{\mathsf{F}} \leq \|\Phi(0)\|_{\mathsf{F}} \lesssim \sqrt{\frac{d}{m}}\|X\|_{\mathsf{F}}$, the penultimate inequality follows from the definition of $\mathrm{sr}(X)$ and the last inequality follows from the lower bound on d_w. Therefore, the sum is bounded by 1, which we use in the second inequality in the following. Putting everything together, we obtain

$$\begin{aligned}\|E(t)\|_{\mathsf{F}} &\leq \frac{2C\eta L^{3/2}\|X\|_{\mathsf{op}}^2\|\Phi(t)\|_{\mathsf{F}}}{m}\sum_{\ell=1}^{L-1}\left(\frac{2C\eta L^{3/2}\|X\|_{\mathsf{op}}\|\Phi(t)\|_{\mathsf{F}}}{\sqrt{md_w}}\right)^{\ell} \\ &\leq \frac{4C^2\eta^2L^3\|X\|_{\mathsf{op}}^3\|\Phi(t)\|_{\mathsf{F}}^2}{m^{3/2}d_w^{1/2}} \cdot \frac{1}{1-\frac{2C\eta L^{3/2}\|X\|_{\mathsf{op}}\|\Phi(t)\|_{\mathsf{F}}}{\sqrt{md_w}}} \\ &\lesssim \frac{\eta^2L^3\|X\|_{\mathsf{op}}^4\|\Phi(t)\|_{\mathsf{F}}}{m^2}\sqrt{\frac{d\,\mathrm{sr}(X)}{d_w}} \\ &\leq \frac{\eta L^2\|X\|_{\mathsf{op}}^2\|\Phi(t)\|_{\mathsf{F}}}{m}\sqrt{\frac{d\,\mathrm{sr}(X)}{d_w}} \\ &\leq \frac{17\eta L\sigma_{\min}^2(X)}{1024m} \cdot \|\Phi(t)\|_{\mathsf{F}},\end{aligned}$$

by using the bound on η and after choosing d_w to satisfy

$$d_w \gtrsim d \cdot \mathrm{sr}(X) \cdot L^2 \cdot \kappa^4.$$

This completes the proof of the Lemma. $\square$

Note that $C_{\mathsf{prod}}^{(i)} \leq 1$ is trivially true. On the other hand, we have the following lower bound.

Lemma A.15. *We have that $C_{\mathsf{prod}}^{(i)} \geq \frac{1}{4}$ for all $1 \leq i \leq L$.*

Proof. From the definition of $C_{\mathsf{prod}}^{(i)}$ in Equation (24) and Theorem B.3, we have $C_{\mathsf{prod}}^{(i)} = \prod_{j\neq i}^{L}\left(1-\frac{\eta\lambda}{d_j}\right) \geq 1-\sum_{j\neq i}^{L}\frac{\eta\lambda}{d_j}$. Moreover, we have that

$$\begin{aligned}
\sum_{j\neq i}\frac{\eta\lambda}{d_j} &\leq \sum_{j\neq i}\frac{m}{d_j}\cdot\frac{\lambda}{L\sigma_{\max}^2(X)} \\
&\leq \frac{\lambda}{L\sigma_{\max}^2(X)} + \sum_{j\notin\{i,1\}}\frac{\lambda}{L\sigma_{\max}^2(X)}\frac{m}{d_w} \\
&\leq \frac{\gamma}{L\kappa^2}\sqrt{\frac{m}{d}}\cdot\left(1+\sum_{j\notin\{i,1\}}\frac{1}{L^2}\right) \\
&\leq \frac{\gamma}{L}\sqrt{\frac{m}{d}}\cdot\left(1+\frac{1}{L}\right) \\
&\leq \frac{3\gamma}{4}\sqrt{\frac{m}{d}} \\
&\leq \frac{3}{4},
\end{aligned}$$

where the first inequality follows from the upper bound on η, the second inequality follows from the fact that $d_j = d_w$ for all $j > 1$ and $d_1 = m$, with $d_w > m$, the third inequality follows from the lower bound $d_w \geq L\cdot m$, the penultimate inequality follows from the fact the function $L \mapsto \frac{1}{L}\left(1+\frac{1}{L}\right)$ is decreasing in L and equal to $\frac{3}{4}$ for $L = 2$, and the last inequality follows from the assumption that $m \leq d$ and $\gamma \leq 1$. ∎

Before we prove the event $\mathcal{A}(t)$, we prove the following Lemma.

Lemma A.16. *Under the event $\mathcal{B}(t)$, the following holds:*

$$\lambda_{\min}(P(t)) \geq \left(\frac{9}{32}\right)^2\cdot\frac{L\sigma_{\min}^2(X)}{m}, \quad \lambda_{\max}(P(t)) \leq \frac{3L\sigma_{\max}^2(X)}{m}.$$

Proof. From Lemma A.15, it follows that $C_{\mathsf{prod}}^{(i)} \in [\frac{1}{4}, 1]$. Consequently Lemma A.7 yields

$$\begin{aligned}
\lambda_{\min}(P(t)) &\geq \frac{1}{4d_w^{L-1}m}\sum_{i=1}^{L}\sigma_{\min}^2(W_{L:(i+1)}(t))\sigma_{\min}^2(W_{(i-1):1}(t)Y) \\
&\geq \frac{1}{4d_w^{L-1}m}\sum_{i=1}^{L}\left(\frac{3}{4}d_w^{\frac{L-i}{2}}\right)^2\left(\frac{3}{4}d_w^{\frac{i-1}{2}}\sigma_{\min}(X)\right)^2 \\
&= \frac{1}{4d_w^{L-1}m}\sum_{i=1}^{L}\left(\frac{9}{16}\right)^2 d_w^{L-1}\sigma_{\min}^2(X) \\
&\geq \frac{81L\sigma_{\min}^2(X)}{1024m},
\end{aligned}$$

where the second inequality uses Equation (59b), the assumption that the event $\mathcal{B}(t)$ holds.

Similarly, we have

$$\lambda_{\max}(P(t)) \leq \frac{1}{d_w^{L-1}m}\sum_{i=1}^{L}\sigma_{\max}^2(W_{L:(i+1)}(t))\sigma_{\max}^2(W_{(i-1):1}(t)Y)$$

$$
\begin{aligned}
&\leq \frac{1}{d_w^{L-1} m} \sum_{i=1}^{L} \left(\frac{5}{4} d_w^{\frac{L-i}{2}}\right)^2 \left(\frac{5}{4} d_w^{\frac{i-1}{2}} \sigma_{\max}(X)\right)^2 \\
&\leq \frac{L\sigma_{\max}^2(X)}{m} \cdot \left(\frac{25}{16}\right)^2 \\
&\leq \frac{3L\sigma_{\max}^2(X)}{m},
\end{aligned}
$$

which completes the proof. □

Lemma A.17. *For any $0 \leq t < \tau$, $\{\{\mathcal{A}(j)\}_{j<t}, \{\mathcal{B}(j)\}_{j\leq t}\} \implies \mathcal{A}(t)$. Moreover, we have*

$$
\mathcal{A}(t) \implies \|\Phi(t+1)\|_{\mathsf{F}} \leq \left(1 - \frac{\eta L \sigma_{\min}^2(X)}{32m}\right) \|\Phi(t)\|_{\mathsf{F}}. \tag{64}
$$

Proof. Recall the decomposition of the error from Equation (39):

$$
\operatorname{vec}(\Phi(t+1)) = (I - \eta P(t))\operatorname{vec}(\Phi(t)) + \operatorname{vec}(E(t)) + (C_{\mathsf{prod}} - 1)\operatorname{vec}(U(t)).
$$

Taking norms on both sides and invoking the bound on $\|E(t)\|_{\mathsf{F}}$ from Lemma A.14, we obtain

$$
\begin{aligned}
\|\operatorname{vec}(\Phi(t+1))\| &= \|(I - \eta P(t))\operatorname{vec}(\Phi(t)) + \operatorname{vec}(E(t)) + (C_{\mathsf{prod}} - 1)\operatorname{vec}(U(t))\| \\
&\leq \|(I - \eta P(t))\|_{\mathsf{op}} \|\Phi(t)\|_{\mathsf{F}} + \|E(t)\|_{\mathsf{F}} + |C_{\mathsf{prod}} - 1| \|U(t)\|_{\mathsf{F}} \\
&\leq \left(1 - \left(\frac{9}{32}\right)^2 \frac{\eta L \sigma_{\min}^2(X)}{m} + \frac{17\eta L \sigma_{\min}^2(X)}{1024m}\right) \|\Phi(t)\|_{\mathsf{F}} + |C_{\mathsf{prod}} - 1| \|U(t)\|_{\mathsf{F}} \\
&\leq \left(1 - \frac{\eta L \sigma_{\min}^2(X)}{16m}\right) \|\Phi(t)\|_{\mathsf{F}} + \left[\frac{(L-1)\eta\lambda}{d_w} + \frac{\eta\lambda}{m}\right] \|U(t)\|_{\mathsf{F}} \\
&\leq \left(1 - \frac{\eta L \sigma_{\min}^2(X)}{16m}\right) \|\Phi(t)\|_{\mathsf{F}} + \frac{5\eta\lambda}{2m} \cdot \|X\|_{\mathsf{F}} \sqrt{\frac{d}{m}},
\end{aligned} \tag{65}
$$

where the penultimate inequality follows from Lemma B.4 and (65) follows from

$$
d_w \geq Lm \implies \frac{(L-1)\eta\lambda}{d_w} \leq \frac{\eta\lambda}{m}, \quad \text{and} \quad \|U(t)\|_{\mathsf{F}} \leq \frac{5}{4} \|X\|_{\mathsf{F}} \sqrt{\frac{d}{m}}
$$

If $t < \tau$, then from the definition of the stopping time τ in (30a) and the identity $\lambda = \gamma\sigma_{\min}^2(X)\sqrt{m/d}$, it follows that

$$
\begin{aligned}
\|\Phi(t+1)\|_{\mathsf{F}} &\leq \left(1 - \frac{\eta L \sigma_{\min}^2(X)}{16m}\right) \|\Phi(t)\|_{\mathsf{F}} + \frac{5\eta\lambda\sqrt{d}}{2m\sqrt{m}} \|X\|_{\mathsf{F}} \\
&\leq \left(1 - \frac{\eta L \sigma_{\min}^2(X)}{32m}\right) \|\Phi(t)\|_{\mathsf{F}},
\end{aligned}
$$

which proves the inequality in (64). □

Corollary A.18. *With high probability, the stopping time τ satisfies*

$$
\tau \leq \frac{32m}{\eta L \sigma_{\min}^2(X)} \log\left(\frac{L\sigma_{\min}^2(X)}{35\lambda}\right).
$$

Proof. For any $t < \tau$, Lemma A.17 implies

$$\begin{aligned}
\|\Phi(t)\|_{\mathsf{F}} &\le \left(1 - \frac{\eta L \sigma_{\min}^2(X)}{32m}\right) \|\Phi(t-1)\|_{\mathsf{F}} \\
&\le \left(1 - \frac{\eta L \sigma_{\min}^2(X)}{32m}\right)^t \|\Phi(0)\|_{\mathsf{F}} \\
&\le \exp\left(-\frac{t\eta L \sigma_{\min}^2(X)}{32m}\right) \|\Phi(0)\|_{\mathsf{F}} \\
&\le \exp\left(-\frac{t\eta L \sigma_{\min}^2(X)}{32m}\right) \frac{11}{5}\sqrt{\frac{d}{m}} \|X\|_{\mathsf{F}},
\end{aligned}$$

where the penultimate inequality follows from the identity $1-x \le \exp(-x)$ and the last inequality follows from Lemma A.11. Finally, we obtain

$$t \ge \frac{32m}{\eta L \sigma_{\min}^2(X)} \log\left(\frac{L\sigma_{\min}^2(X)}{35\lambda}\right) \implies \|\Phi(t)\|_{\mathsf{F}} \le \frac{80\lambda \|X\|_{\mathsf{F}}}{L\sigma_{\min}^2(X)}\sqrt{\frac{d}{m}} = \frac{80\gamma\|X\|_{\mathsf{F}}}{L},$$

which implies the stated upper bound on τ. □

Next we will prove the event $\mathcal{C}(t)$ (Equation (59c)).

Lemma A.19. *For any $t \le \tau$, we have that $\{\mathcal{A}(j), \mathcal{B}(j)\}_{j<t} \implies \mathcal{C}(t)$:*

$$\|W_i(t) - \left(1 - \frac{\eta\lambda}{d_i}\right)^t W_i(0)\|_{\mathsf{F}} \lesssim \frac{\kappa^2\sqrt{d\,\mathrm{sr}(X)}}{L} := R, \quad \textit{for all } i = 1, \ldots, L. \tag{66}$$

Proof. Given Lemma A.5 we obtain the bound

$$\begin{aligned}
&\|W_i(t) - \left(1 - \frac{\eta\lambda}{d_i}\right)^t W_i(0)\|_{\mathsf{F}} \\
&\le \eta \sum_{j=0}^{t-1} \left(1 - \frac{\eta\lambda}{d_i}\right)^{t-1-j} d_w^{-\frac{L-1}{2}} m^{-\frac{1}{2}} \|W_{L:(i+1)}(j)\Phi(j)W_{(i-1):1}(j)Y\|_{\mathsf{F}} \\
&\le \eta \sum_{j=0}^{t-1} \left(1 - \frac{\eta\lambda}{d_i}\right)^{t-1-j} d_w^{-\frac{L-1}{2}} m^{-\frac{1}{2}} \|W_{L:(i+1)}(j)\|_{\mathsf{op}} \|\Phi(j)\|_{\mathsf{F}} \|W_{(i-1):1}(j)Y\|_{\mathsf{op}} \\
&\le \eta \sum_{j=0}^{t-1} \left(1 - \frac{\eta\lambda}{d_i}\right)^{t-1-j} \left(1 - \frac{\eta L \sigma_{\min}^2(X)}{32m}\right)^j \left(\frac{5}{4} d_w^{\frac{L-i}{2}}\right) \left(\frac{5}{4} d_w^{\frac{i-1}{2}} \sigma_{\max}(X)\right) \frac{\|\Phi(0)\|_{\mathsf{F}}}{d_w^{\frac{L-1}{2}} m^{\frac{1}{2}}} \\
&\le \eta \frac{25\|\Phi(0)\|_{\mathsf{F}}\|X\|_{\mathsf{op}}}{16\sqrt{m}} \sum_{j=0}^{t-1} \left(1 - \frac{\eta L \sigma_{\min}^2(X)}{32m}\right)^j \\
&\le \eta \frac{25\|\Phi(0)\|_{\mathsf{F}}\|X\|_{\mathsf{op}}}{16\sqrt{m}} \cdot \frac{32m}{\eta L \sigma_{\min}^2(X)} \\
&= \frac{C\|\Phi(0)\|_{\mathsf{F}}\|X\|_{\mathsf{op}}\sqrt{m}}{L\sigma_{\min}^2(X)} \\
&\le \frac{C\sqrt{d} \cdot \|X\|_{\mathsf{F}}\|X\|_{\mathsf{op}}}{L\sigma_{\min}^2(X)}
\end{aligned}$$

$$\leq \frac{\kappa^2 \sqrt{d\,\mathrm{sr}(X)}}{L},$$

with $C = 50$, which is independent of the choice of layer i. This completes the proof. □

Finally we prove the event $\mathcal{B}(t)$ (Equation (59b)).

Lemma A.20. *We have that $\mathcal{C}(t) \implies \mathcal{B}(t)$ for any $t \leq \tau$.*

Proof. To prove $\mathcal{B}(t)$, we need to control the extremal singular values of several matrix products.

Bounding $\|W_{j:i}(t)\|_{\mathsf{op}}$. Fix any $i > 1$ and $j \geq i$. We start with the following decomposition:

$$\begin{aligned}
W_{j:i}(t) &= \prod_{\ell=j}^{i} W_\ell(t) \\
&= \prod_{\ell=j}^{i} \left(\left(1 - \frac{\eta\lambda}{d_\ell}\right)^t W_\ell(0) + \underbrace{W_\ell(t) - \left(1 - \frac{\eta\lambda}{d_\ell}\right)^t W_\ell(0)}_{\Delta_\ell(t)} \right) \\
&= W_{j:i}(0) \cdot \prod_{\ell=j}^{i} \left(1 - \frac{\eta\lambda}{d_\ell}\right)^t + \sum_{s=1}^{j-i+1} \sum_{i \leq k_1, \dots, k_s \leq j} \widetilde{W}_{j:(k_s+1)}(0) \Delta_{k_s}(t) \dots \Delta_{k_1}(t) \widetilde{W}_{(k_1-1):i}(0),
\end{aligned}$$

using a slight abuse of the notation for $\widetilde{W}_{j:i}$ introduced in (A.5):

$$\widetilde{W}_{j:i}(0) = W_{j:i} \cdot \prod_{\ell=j}^{i} \left(1 - \frac{\eta\lambda}{d_\ell}\right)^t.$$

Continuing, we have the following upper bound:

$$\begin{aligned}
\|\widetilde{W}_{j:(k_s+1)}(0) \Delta_{k_s}(t) \dots \Delta_{k_1}(t) \widetilde{W}_{(k_1-1):i}(0)\|_{\mathsf{op}} &\leq R^s \left[\prod_{\substack{\ell=j \\ \ell \notin \{k_1, \dots, k_s\}}}^{i} \left(1 - \frac{\eta\lambda}{d_\ell}\right)^t \right] \left(\sqrt{\frac{L}{c}}\right)^{s+1} d_w^{\frac{j-i+1-s}{2}} \\
&\leq \sqrt{\frac{L}{c}} d_w^{\frac{j-i+1}{2}} \cdot \left(\frac{CR\sqrt{L}}{\sqrt{d_w}} \right)^s
\end{aligned}$$

Summing up over all possible $k_1, \dots, k_s$ for all possible $s = 1$ to $s = j - i + 1$, we have

$$\begin{aligned}
&\sum_{s=1}^{j-i+1} \sum_{i \leq k_1, \dots, k_s \leq j} \|\widetilde{W}_{j:(k_s+1)}(0) \Delta_{k_s}(t) \dots \Delta_{k_1}(t) \widetilde{W}_{(k_1-1):i}(0)\|_{\mathsf{op}} \\
&\leq \sqrt{\frac{L}{c}} d_w^{\frac{j-i+1}{2}} \sum_{s=1}^{j-i+1} \binom{j-i+1}{s} \cdot \left(\frac{CR\sqrt{L}}{\sqrt{d_w}} \right)^s \\
&\leq \sqrt{\frac{L}{c}} d_w^{\frac{j-i+1}{2}} \sum_{s=1}^{j-i+1} \left(\frac{CR\sqrt{L}}{\sqrt{d_w}} \right)^s,
\end{aligned}$$

where the last inequality follows from the bound $\binom{j-i+1}{s} \leq \binom{L}{s} \leq L^s$. Finally, by Lemma B.5,

$$\sum_{s=1}^{j-i+1} \left(\frac{CRL^{3/2}}{\sqrt{d_w}}\right)^s \lesssim \left(\frac{CRL^{3/2}}{\sqrt{d_w}}\right) \cdot \frac{1}{1 - \left(\frac{CRL^{3/2}}{\sqrt{d_w}}\right)} \leq \frac{1}{3},$$

as long as we choose d_w such that

$$d_w \gtrsim R^2 L^3 \asymp \frac{Ld\kappa^4}{\sigma_{\max}^2(X)} \Leftrightarrow \left(\frac{RL^{3/2}}{\sqrt{d_w}}\right) \lesssim \frac{1}{4}.$$

Putting everything together, we arrive at

$$\|W_{j:i}(t)\|_{\mathsf{op}} \leq \|W_{j:i}(0)\|_{\mathsf{op}} \prod_{\ell=j}^{i} \left(1 - \frac{\eta\lambda}{d_\ell}\right)^t + \frac{1}{3}\sqrt{\frac{L}{c}} d_w^{\frac{j-i+1}{2}} \leq \frac{4}{3}\sqrt{\frac{L}{c}} d_w^{\frac{j-i+1}{2}},$$

using $(1 - \eta\lambda/d_\ell) \leq 1$ and Lemma A.9 in the last inequality. This proves the first bound in the definition of $\mathcal{B}(t)$.

Bounding $\sigma_{\min}(W_{i:1}Y)$ and $\|W_{i:1}Y\|_{\mathsf{op}}$. To control the singular values of $W_{i:1}(t)Y$ for $i < L$, we write

$$\begin{aligned}
W_{i:1}(t)Y &= \Big(\prod_{\ell=i}^{1} W_\ell(t)\Big)Y \\
&= \left[\prod_{\ell=i}^{1} \left(\left(1 - \frac{\eta\lambda}{d_\ell}\right)^t W_\ell(0) + \left[W_\ell(t) - \left(1 - \frac{\eta\lambda}{d_\ell}\right)^t W_\ell(0)\right]\right)\right]Y \\
&= W_{i:1}(0)Y \cdot \prod_{\ell=i}^{1} \left(1 - \frac{\eta\lambda}{d_\ell}\right)^t + \sum_{s=1}^{i} \sum_{1 \leq k_1, \dots, k_s \leq i} \widetilde{W}_{i:(k_s+1)}(0)\Delta_{k_s} \dots \Delta_{k_1} \widetilde{W}_{(k_1-1):1}(0)Y
\end{aligned}$$

From the above decomposition and Weyl's inequality, it follows that

$$|\sigma_j(W_{i:1}(t)Y) - \sigma_j(\widetilde{W}_{i:1}(0)Y)| \leq \sum_{s=1}^{i} \sum_{(k_1,\dots,k_s)} \|\widetilde{W}_{i:(k_s+1)}(0)\Delta_{k_s} \dots \Delta_{k_1} \widetilde{W}_{(k_1-1):1}(0)Y\|_{\mathsf{op}}. \tag{67}$$

We now turn to bound the terms in the sum on the RHS of (67). First, note that

$$\begin{aligned}
&\|\widetilde{W}_{i:(k_s+1)}(0)\Delta_{k_s} \dots \Delta_{k_1} \widetilde{W}_{(k_1-1):1}(0)Y\|_{\mathsf{op}} \\
&= \prod_{\substack{\ell=i \\ \ell \notin \{k_1,\dots,k_s\}}}^{1} \left(1 - \frac{\eta\lambda}{d_\ell}\right)^t \|W_{i:(k_s+1)}(0)\Delta_{k_s} \dots \Delta_{k_1} W_{(k_1-1):1}(0)Y\|_{\mathsf{op}} \\
&\leq \|W_{i:(k_s+1)}(0)\Delta_{k_s} \dots \Delta_{k_1} W_{(k_1-1):1}(0)Y\|_{\mathsf{op}} \\
&\leq \left(R \cdot 2\sqrt{\frac{L}{c}}\right)^s d_w^{\frac{i-k_1+1-s}{2}} \cdot \frac{6}{5} d_w^{\frac{k_1-1}{2}} \sigma_{\max}(X)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{6}{5} \cdot \left(R \cdot 2\sqrt{\frac{L}{c}} \right)^s d_w^{\frac{i-s}{2}} \sigma_{\max}(X) \\
&= \frac{6}{5} \left(R \cdot 2\sqrt{\frac{L}{cd_w}} \right)^s d_w^{\frac{i}{2}} \cdot \sigma_{\max}(X).
\end{aligned}
$$

Again, summing over all possible $(k_1, \ldots, k_s)$ for $s = 1$ to i yields the upper bound

$$
\begin{aligned}
\frac{6\sigma_{\max}(X) \cdot d_w^{\frac{i}{2}}}{5} \cdot \sum_{s=1}^{i} \binom{i}{s} \left(R \cdot 2\sqrt{\frac{L}{cd_w}} \right)^s &\leq \frac{6\sigma_{\max}(X) \cdot d_w^{\frac{i}{2}}}{5} \sum_{s=1}^{i} \left(R \cdot 2i\sqrt{\frac{L}{cd_w}} \right)^s \\
&\leq \frac{6\sigma_{\max}(X) \cdot d_w^{\frac{i}{2}}}{5} \frac{R \cdot 2L\sqrt{\frac{L}{cd_w}}}{1 - R \cdot 2L\sqrt{\frac{L}{cd_w}}} \\
&\leq \frac{6\sigma_{\max}(X) \cdot d_w^{\frac{i}{2}}}{5} \frac{R \cdot 4L\sqrt{\frac{L}{cd_w}}}{3} \\
&\leq c_{\mathsf{b}} \cdot \sigma_{\min}(X) \cdot d_w^{\frac{i}{2}},
\end{aligned}
$$

valid for any d_w satisfying the following identity:

$$
\frac{R \cdot 8L}{5} \sqrt{\frac{L}{cd_w}} \leq \frac{c_{\mathsf{b}}}{\kappa} \Leftrightarrow d_w \gtrsim \kappa^2 R^2 L^3 c_{\mathsf{b}}^2 \asymp Ld \cdot \frac{\kappa^6 \operatorname{sr}(X)}{c_{\mathsf{b}}^2},
$$

where c_{b} is a free parameter. Plugging the derived bound into (67) yields

$$
\begin{aligned}
\sigma_{\max}(W_{i:1}(t)Y) &\leq \sigma_{\max}(\widetilde{W}_{i:1}(0)Y) + c_{\mathsf{b}} \cdot \sigma_{\min}(X) \cdot d_w^{\frac{i}{2}} \\
&\leq \sigma_{\max}(W_{i:1}(0)Y) + c_{\mathsf{b}} \cdot \sigma_{\max}(X) \cdot d_w^{\frac{i}{2}} \\
&\leq \left(\frac{6}{5} + c_{\mathsf{b}} \right) \sigma_{\max}(X) \cdot d_w^{\frac{i}{2}} \\
&\leq \frac{5}{4} \sigma_{\max}(X) \cdot d_w^{\frac{i}{2}},
\end{aligned}
$$

after choosing $c_{\mathsf{b}} \leq \frac{1}{20}$. This proves the second bound in the definition of $\mathcal{B}(t)$.

Similarly, we have the following lower bound:

$$
\begin{aligned}
\sigma_{\min}(W_{i:1}(t)Y) &\geq \sigma_{\min}(\widetilde{W}_{i:1}(0)Y) - c_{\mathsf{b}} \cdot \sigma_{\min}(X) \cdot d_w^{\frac{i}{2}} \\
&\geq \sigma_{\min}(W_{i:1}(0)Y) \cdot \prod_{\ell=i}^{1} \left(1 - \frac{\eta\lambda}{d_\ell} \right)^t - c_{\mathsf{b}} \cdot \sigma_{\min}(X) \cdot d_w^{\frac{i}{2}} \\
&\geq \left[\left(1 - \frac{1}{20L} \right)^i \cdot \frac{4}{5} - c_{\mathsf{b}} \right] \cdot \sigma_{\min}(X) \cdot d_w^{\frac{i}{2}} \\
&\geq \left[\left(1 - \frac{1}{20L} \right)^L \cdot \frac{4}{5} - c_{\mathsf{b}} \right] \cdot \sigma_{\min}(X) \cdot d_w^{\frac{i}{2}} \\
&\geq \left[\frac{19}{20} \cdot \frac{4}{5} - c_{\mathsf{b}} \right] \cdot \sigma_{\min}(X) \cdot d_w^{\frac{i}{2}}
\end{aligned}
$$

$$\geq \frac{3\sigma_{\min}(X)}{4} \cdot d_w^{\frac{i}{2}},$$

where the third inequality follows from Lemma A.21, the second to last inequality follows from Lemma A.22 and the last inequality follows from choosing $c_{\mathsf{b}} \leq \frac{1}{100}$. This proves the fourth bound in the definition of $\mathcal{B}(t)$.

Bounding $\sigma_{\min}(W_{L:i})$ and $\|W_{L:i}\|_{\mathsf{op}}$. We now furnish upper and lower bounds for singular values of $W_{L:i}(t)$, for $i > 1$. By an analogous argument to the one employed for $W_{j:i}(t)$, when $j < L$, we arrive at

$$W_{L:i}(t) = W_{L:i}(0) \prod_{\ell=L}^{i} \left(1 - \frac{\eta\lambda}{d_\ell}\right)^t + \sum_{s=1}^{L-i} \sum_{i \leq k_1, \dots, k_s \leq L} \widetilde{W}_{L:(k_s+1)}(0)\Delta_{k_s} \dots \Delta_{k_1} \widetilde{W}_{(k_1-1):i}(0) \tag{68}$$

As before, we bound each summand on the RHS of (68). We have

$$\begin{aligned}
\|\widetilde{W}_{L:(k_s+1)}(0)\Delta_{k_s} \dots \Delta_{k_1} \widetilde{W}_{(k_1-1):i}(0)\|_{\mathsf{op}} &\leq R^s \cdot \frac{6}{5} d_w^{\frac{L-k_s}{2}} \cdot \left(2\sqrt{\frac{L}{c}}\right)^s d_w^{\frac{k_s-i+1-s}{2}} \\
&= \frac{6}{5} \cdot \left(\frac{RL^{1/2}}{\sqrt{d_w}}\right)^s \cdot d_w^{\frac{L-i+1}{2}}
\end{aligned}$$

Adding up all the summands yields the upper bound

$$\begin{aligned}
\sum_{s=1}^{L-i} \sum_{i \leq k_1, \dots, k_s \leq L} \|\widetilde{W}_{L:(k_s+1)}(0)\Delta_{k_s} \dots \Delta_{k_1} \widetilde{W}_{(k_1-1):i}(0)\|_{\mathsf{op}} &\leq \frac{6 d_w^{\frac{L-i+1}{2}}}{5} \cdot \sum_{s=1}^{L-i} \binom{L-i}{s} \left(\frac{RL^{1/2}}{\sqrt{d_w}}\right)^s \\
&\leq \frac{6 d_w^{\frac{L-i+1}{2}}}{5} \cdot \sum_{s=1}^{L-i} \left(\frac{RL^{3/2}}{\sqrt{d_w}}\right)^s \\
&\leq \frac{6 d_w^{\frac{L-i+1}{2}}}{5} \cdot \frac{RL^{3/2}}{\sqrt{d_w}},
\end{aligned}$$

where the penultimate inequality follows from $\binom{k}{i} \leq k^i$ and the last inequality follows from Lemma B.5. Again, we introduce a free parameter $\bar{c}_{\mathsf{b}}$ and require

$$\frac{RL^{3/2}}{\sqrt{d_w}} = \frac{\kappa^2 \sqrt{Ld\,\mathrm{sr}(X)}}{\sqrt{d_w}} \lesssim \bar{c}_{\mathsf{b}} \Leftrightarrow d_w \gtrsim \frac{Ld\,\mathrm{sr}(X)\kappa^4}{\bar{c}_{\mathsf{b}}^2}.$$

Returning to (68), we obtain the upper bound

$$\sigma_{\max}(W_{L:i}(t)) \leq \sigma_{\max}(W_{L:i}(0)) + \bar{c}_{\mathsf{b}} \cdot d_w^{\frac{L-i+1}{2}} \leq \left(\frac{6}{5} + \bar{c}_{\mathsf{b}}\right) \cdot d_w^{\frac{L-i+1}{2}} \leq \frac{5}{4} \cdot d_w^{\frac{L-i+1}{2}},$$

choosing $\bar{c}_{\mathsf{b}} \leq \frac{1}{20}$. This proves the third inequality in $\mathcal{B}(t)$; similarly by using Lemma A.21 and Lemma A.22 we get the lower bound

$$\sigma_{\min}(W_{L:i}(t)) \geq \sigma_{\min}(W_{L:i}(0)) \cdot \left(1 - \frac{1}{20L}\right)^{L-i+1} - \bar{c}_{\mathsf{b}} \cdot d_w^{\frac{L-i+1}{2}}$$

$$
\begin{aligned}
&\geq \sigma_{\min}(W_{L:i}(0)) \cdot \left(1 - \frac{1}{20L}\right)^{L} - \bar{c}_{\mathsf{b}} \cdot d_w^{\frac{L-i+1}{2}} \\
&\geq \left(0.95 \cdot \frac{4}{5} - \bar{c}_{\mathsf{b}}\right) \cdot d_w^{\frac{L-i+1}{2}} \\
&\geq \frac{3}{4} \cdot d_w^{\frac{L-i+1}{2}},
\end{aligned}
$$

after choosing $\bar{c}_{\mathsf{b}} \leq \frac{1}{100}$. This proves the final inequality making up the event $\mathcal{B}(t)$. $\square$

Lemma A.21. *For any $t \leq \tau$, it follows that*

$$\left(1 - \frac{\eta\lambda}{d_i}\right)^t \geq 1 - \frac{1}{20L}.$$

Proof. From Theorem B.3, it follows that

$$\left(1 - \frac{\eta\lambda}{d_i}\right)^t \geq 1 - \frac{t\eta\lambda}{d_i} \geq 1 - \frac{\tau\eta\lambda}{d_i}.$$

From Corollary A.18, the quantity above is at least

$$
\begin{aligned}
1 - \frac{\tau\eta\lambda}{d_i} &\geq 1 - \frac{32m}{\eta\sigma_{\min}^2(X)L} \log\left(\frac{L\sigma_{\min}^2(X)}{35\lambda}\right) \cdot \frac{\eta\lambda}{d_i} \\
&= 1 - 32 \cdot \frac{\lambda}{L\sigma_{\min}^2(X)} \cdot \frac{m}{d_i} \cdot \log\left(\frac{L\sigma_{\min}^2(X)}{35\lambda}\right).
\end{aligned}
$$

We now argue that for small enough λ, the last term is at most $1 - \frac{1}{20L}$. Indeed,

$$\frac{35\lambda}{L\sigma_{\min}^2(X)} \log\left(\frac{L\sigma_{\min}^2(X)}{35\lambda}\right) \leq \frac{1}{20} \Leftrightarrow 20 \leq \frac{20 \cdot 35\lambda}{L\sigma_{\min}^2(X)} \exp\left(\frac{L\sigma_{\min}^2(X)}{20 \cdot 35\lambda}\right).$$

The above inequality is itself implied by the assumption that $\lambda \leq \frac{L\sigma_{\min}^2(X)}{400 \cdot 35}$, which implies

$$\frac{20 \cdot 35\lambda}{L\sigma_{\min}^2(X)} \exp\left(\frac{L\sigma_{\min}^2(X)}{20 \cdot 35\lambda}\right) \geq \frac{L\sigma_{\min}^2(X)}{20 \cdot 35\lambda} \geq 20,$$

using the inequality $x\exp(1/x) \geq \frac{1}{x}$ for all $x > 0$ above. Therefore,

$$1 - \frac{\tau\eta\lambda}{d_i} \geq 1 - \frac{32}{35}\frac{35\lambda}{L\sigma_{\min}^2(X)} \log\left(\frac{L\sigma_{\min}^2(X)}{35\lambda}\right) \frac{m}{d_i} \geq 1 - \frac{32}{35 \cdot 20L} \geq 1 - \frac{1}{20L},$$

which completes the proof of the claim. $\square$

Lemma A.22. *For any $L \geq 2$, we have that*

$$\left(1 - \frac{1}{20L}\right)^L \geq 0.95.$$

Proof. The function $x \mapsto \left(1 - \frac{1}{20x}\right)^x$ is monotone increasing for all $x \geq 1$. Therefore,

$$\left(1 - \frac{1}{20L}\right)^L \geq \left(1 - \frac{1}{40}\right)^2 = \left(\frac{39}{40}\right)^2 > 0.95.$$

$\square$

Proof of Theorem A.12. We prove this theorem by induction. The base case follows from Lemma A.13. Now, suppose that all events $\mathcal{A}(t)$, $\mathcal{B}(t)$ and $\mathcal{C}(t)$ hold up to some arbitrary index $t < \tau$. Then:

- The event $\mathcal{C}(t+1)$ holds by Lemma A.19;
- The event $\mathcal{B}(t+1)$ holds by Lemma A.20 and the previous item;
- Finally, the event $\mathcal{A}(t+1)$ holds by Lemma A.17 and the preceding item.

This completes the proof of the theorem. □

A.6 Step 2: Regression error stays small

In Appendix A.5 we have shown that after τ iterations our regression error is small; namely, $\|\Phi(\tau)\|_{\mathsf{F}} \leq \frac{80\gamma}{L}\|X\|_{\mathsf{F}}$. We now want to show that the regression error remains small until at least iteration T. In particular, we will show that $\|\Phi(t)\|_{\mathsf{F}} \leq C_1\gamma\|X\|_{\mathsf{F}}$ for all $\tau \leq t \leq T$. This we will show again by induction over the events stated in the following theorem.

Theorem A.23. *Given τ defined in (30a) and T defined in (30b), then for all $\tau \leq t \leq T$ the following events hold with probability of at least $1 - e^{-\Omega(d)}$ over the random initialization,*

$$\mathcal{A}(t) := \{\|\Phi(t)\|_{\mathsf{F}} \leq C_1\gamma\|X\|_{\mathsf{F}}\} \tag{69a}$$

$$\mathcal{B}(t) := \left\{ \begin{array}{rcl} \sigma_{\max}(W_{j:i}(t)) & \leq & \left(2\sqrt{\frac{L}{c}}\right) d_w^{\frac{j-i+1}{2}}, \ \ \forall 1 < i \leq j < L \\ \sigma_{\max}(W_{i:1}(t)Y) & \leq & \frac{9}{7} d_w^{\frac{i}{2}} \sigma_{\max}(X), \ \ \forall 1 \leq i < L \\ \sigma_{\max}(W_{L:i}(t)) & \leq & \frac{9}{7} d_w^{\frac{L-i+1}{2}}, \ \ \forall 1 < i \leq L \\ \sigma_{\min}(W_{L:i}(t)) & \geq & \frac{5}{7} d_w^{\frac{L-i+1}{2}}, \ \ \forall 1 < i \leq L \end{array} \right\} \tag{69b}$$

$$\mathcal{C}(t) := \left\{ \|W_i(t) - \left(1 - \frac{\eta\lambda}{d_i}\right)^{t-\tau} W_i(\tau)\|_{\mathsf{F}} \leq \Delta_\infty \right\}, \quad \Delta_\infty := C\kappa^2\sqrt{d\,\mathrm{sr}(X)}\log(d_w). \tag{69c}$$

where $C_1 > 0$ is a universal constant and $c > 0$ is the constant from Lemma A.9.

The events are similar to those in the first phase. In this phase, the difference is that we cannot guarantee anymore that the smallest singular value of $\sigma_{\min}(W_{i:1}Y)$ gets arbitrarily small. Note that $\mathcal{A}(\tau), \mathcal{B}(\tau)$ are true by Theorem A.12 and $\mathcal{C}(\tau)$ is trivially true. Throughout this section, we will require d_w to satisfy the following inequality

$$d_w \gtrsim \Delta_\infty^2 L^3 = \mathcal{O}\left(L^3\kappa^4 d\,\mathrm{sr}(X)\log^2(d_w)\right). \tag{70}$$

Proof of $\mathcal{C}(t)$. We start by proving $\mathcal{C}(t)$ given $\{\mathcal{A}(j), \mathcal{B}(j)\}_{j<t}$.

Lemma A.24. *Given that the set of events $\{\mathcal{A}(j), \mathcal{B}(j)\}_{j=\tau}^{t-1}$ for $\tau \leq t \leq T$ hold, then $\mathcal{C}(t)$ holds:*

$$\|W_i(t) - \left(1 - \frac{\eta\lambda}{d_i}\right)^{t-\tau} W_i(\tau)\|_{\mathsf{F}} \lesssim \kappa^2\sqrt{d\,\mathrm{sr}(X)}\log(d_w).$$

Proof. From Lemma A.5 and the trivial bound $(1 - \eta\lambda/d_i) \leq 1$, we deduce that

$$\|W_i(t) - \left(1 - \frac{\eta\lambda}{d_i}\right)^{t-\tau} W_i(\tau)\|_{\mathsf{F}}$$

$$
\begin{aligned}
&\le \eta \sum_{j=0}^{t-\tau-1} \left(1-\frac{\eta\lambda}{d_i}\right)^{t-\tau-j} \frac{1}{\sqrt{d_w^{L-1}m}} \|W_{L:(i+1)}(j+\tau)\|_{\mathsf{op}} \|\Phi(j+\tau)\|_{\mathsf{F}} \|W_{(i-1):1}(j+\tau)Y\|_{\mathsf{op}} \\
&\le \frac{\eta}{\sqrt{d_w^{L-1}m}} \sum_{j=0}^{t-\tau-1} \frac{9}{7} d_w^{\frac{L-i}{2}} \cdot \|\Phi(j+\tau)\|_{\mathsf{F}} \cdot \frac{9}{7} d_w^{\frac{i-1}{2}} \sigma_{\max}(X) \\
&\lesssim \frac{\eta\|X\|_{\mathsf{op}}}{\sqrt{m}} \sum_{j=0}^{t-\tau-1} \|\Phi(j+\tau)\|_{\mathsf{F}} \\
&\lesssim \frac{\eta\|X\|_{\mathsf{op}}}{\sqrt{m}} \cdot T\gamma\|X\|_{\mathsf{F}} \\
&\lesssim \kappa^2 \sqrt{d\,\mathrm{sr}(X)} \cdot \log(d_w),
\end{aligned}
$$

where the second inequality follows from $\{\mathcal{B}(j)\}_{j=\tau}^{t-1}$, the fourth inequality follows from $\{\mathcal{A}(j)\}_{j=\tau}^{t-1}$, and the last inequality follows from the upper bound on η and the identity $\|X\|_{\mathsf{F}}\|X\|_{\mathsf{op}} = \|X\|_{\mathsf{op}}^2\sqrt{\mathrm{sr}(X)}$. $\square$

We next prove that $\mathcal{B}(t)$ is implied by $\mathcal{C}(t)$.

Lemma A.25. *Fix $t \in [\tau, T]$. Then $\mathcal{C}(t) \implies \mathcal{B}(t)$.*

Proof. As in the proof of Lemma A.20, we have the decomposition

$$
\begin{aligned}
W_{j:i}(t) &= \prod_{\ell=i}^{j} \left(\left(1-\frac{\eta\lambda}{d_\ell}\right)^{t-\tau} W_\ell^{(\tau)} + \left(W_\ell^{(t)} - \left(1-\frac{\eta\lambda}{d_\ell}\right)^{t-\tau} W_\ell^{(\tau)} \right) \right) \\
&= W_{j:i}(\tau) \prod_{\ell=j}^{i} \left(1-\frac{\eta\lambda}{d_\ell}\right)^{t-\tau} \\
&\quad + \sum_{i \le k_1, \dots, k_s \le j} \left[\prod_{\ell \notin \{k_1,\dots,k_s\}} \left(1-\frac{\eta\lambda}{d_\ell}\right)^{t-\tau} \right] W_{j:(k_s+1)}(\tau)\Delta_{k_s} \dots \Delta_{k_1} W_{(k_1-1):i}(\tau),
\end{aligned}
$$

where each term satisfies $\|\Delta_{k_i}\|_{\mathsf{op}} \le \Delta_\infty$. Therefore,

$$
\begin{aligned}
\left\| W_{j:i}(t) - W_{j:i}(\tau) \prod_{\ell=j}^{i} \left(1-\frac{\eta\lambda}{d_\ell}\right)^{t-\tau} \right\|_{\mathsf{op}} &\le \sum_{s=1}^{j-i+1} \binom{j-i+1}{s} (\Delta_\infty)^s \left(2\sqrt{\frac{L}{c}}\right)^{s+1} d_w^{\frac{j-i+1-s}{2}} \\
&\le \left(2\sqrt{\frac{L}{c}}\right) d_w^{\frac{j-i+1}{2}} \sum_{s=1}^{j-i} \left(\frac{C\Delta_\infty L^{3/2}}{\sqrt{d_w}}\right)^s \\
&\le \left(2\sqrt{\frac{L}{c}}\right) d_w^{\frac{j-i+1}{2}} \frac{1}{31},
\end{aligned} \tag{71}
$$

using Lemma B.5 and the assumed bound (70). Therefore,

$$
\|W_{j:i}(t)\|_{\mathsf{op}} \le \|W_{j:i}(\tau)\|_{\mathsf{op}} \prod_{\ell=j}^{i} \left(1-\frac{\eta\lambda}{d_\ell}\right)^{t-\tau} + \left\| W_{j:i}(t) - W_{j:i}(\tau) \prod_{i \le \ell \le j} \left(1-\frac{\eta\lambda}{d_\ell}\right)^{t-\tau} \right\|_{\mathsf{op}}
$$

$$
\begin{aligned}
&\le \left(2\sqrt{\frac{L}{c}}\right) d_w^{\frac{j-i+1}{2}} + \left(2\sqrt{\frac{L}{c}}\right) d_w^{\frac{j-i+1}{2}} \frac{1}{31} \\
&\le \left(2\sqrt{\frac{L}{c}}\right) d_w^{\frac{j-i+1}{2}},
\end{aligned}
$$

relabeling c appropriately in the last step to absorb the $1 + \frac{1}{31}$ term. This proves the first bound in the event $\mathcal{B}(t)$. Continuing with $W_{L:i}(t)$, we have

$$
\begin{aligned}
\left\| W_{L:i}(t) - W_{L:i}(\tau) \prod_{i \le \ell \le L} \left(1 - \frac{\eta\lambda}{d_\ell}\right)^{t-\tau} \right\|_{\mathsf{op}} &\le \sum_{s=1}^{L-i+1} \binom{i}{\ell} (\Delta_\infty)^s \left(2\sqrt{\frac{L}{c}}\right)^s d_w^{\frac{L-i+1-s}{2}} \frac{5}{4} \\
&\le \frac{5}{4} d_w^{\frac{L-i+1}{2}} \sum_{s=1}^{L-i+1} \left(\frac{C\Delta_\infty L^{3/2}}{\sqrt{d_w}}\right)^s \\
&\le \frac{1}{63}\frac{5}{4} d_w^{\frac{L-i+1}{2}}. \qquad (72)
\end{aligned}
$$

Again using the bound from (70), we deduce that

$$
\begin{aligned}
\|W_{L:i}(t)\|_{\mathsf{op}} &\le \|W_{L:i}(\tau)\|_{\mathsf{op}} + \left\| W_{L:i}(t) - W_{L:i}(\tau) \prod_{i \le \ell \le L} \left(1 - \frac{\eta\lambda}{d_\ell}\right)^{t-\tau} \right\|_{\mathsf{op}} \\
&\le \frac{5}{4} d_w^{\frac{L-i+1}{2}} + \frac{1}{63}\frac{5}{4} d_w^{\frac{L-i+1}{2}} \\
&= \frac{80}{63} d_w^{\frac{L-i+1}{2}} \\
&\le \frac{9}{7} d_w^{\frac{L-i+1}{2}},
\end{aligned}
$$

which proves the second bound from the event $\mathcal{B}(t)$, as well as

$$
\begin{aligned}
\sigma_{\min}(W_{L:i}(t)) &\ge \sigma_{\min}(W_{L:i}(\tau)) \prod_{i \le \ell \le L} \left(1 - \frac{\eta\lambda}{d_\ell}\right)^{t-\tau} - \left\| W_{L:i}(t) - W_{L:i}(\tau) \prod_{i \le \ell \le L} \left(1 - \frac{\eta\lambda}{d_\ell}\right)^{t-\tau} \right\|_{\mathsf{op}} \\
&\ge \frac{3}{4} d_w^{\frac{L-i+1}{2}} \cdot \prod_{i \le \ell \le L} \left(1 - \frac{\eta\lambda}{d_\ell}\right)^{t-\tau} - \frac{1}{63}\frac{3}{4} d_w^{\frac{L-i+1}{2}} \\
&\ge \frac{3}{4} d_w^{\frac{L-i+1}{2}} \left[\exp\left(-2 \cdot \frac{(t-\tau)\eta\lambda}{d_\ell}\right)^{L-i+1} - \frac{1}{63} \right] \\
&\ge \frac{3}{4} d_w^{\frac{L-i+1}{2}} \left[\exp\left(-2 \cdot \frac{L \cdot \log(d_w) \cdot m}{d_\ell}\right) - \frac{1}{63} \right] \\
&\ge \frac{3}{4} d_w^{\frac{L-i+1}{2}} \cdot \frac{60}{63} \\
&= \frac{5}{7} d_w^{\frac{L-i+1}{2}},
\end{aligned}
$$

using $1 - x \ge \exp(-2x)$ in the third inequality, $t - \tau \le {}^{\log(d_w) \cdot m}\!/\!{}_{\eta\lambda}$ in the penultimate inequality, the fact that $d_\ell = d_w$ for $\ell > 1$, and choosing $d_w \ge {}^{4L\log(d_w)\cdot m}\!/\!{}_{\log(\frac{63}{61})}$ in the last inequality. This proves the third bound from $\mathcal{B}(t)$.

Finally, we have the upper bound

$$
\begin{aligned}
\|W_{i:1}(t)Y\|_{\mathsf{op}} &\leq \|W_{i:1}(\tau)Y\|_{\mathsf{op}} \prod_{1\leq \ell\leq i}\left(1-\frac{\eta\lambda}{d_\ell}\right)^{t-\tau} + \left\|W_{i:1}(t)Y - W_{i:1}(\tau)Y\prod_{1\leq \ell\leq i}\left(1-\frac{\eta\lambda}{d_\ell}\right)^{t-\tau}\right\|_{\mathsf{op}} \\
&\leq \frac{5}{4}d_w^{\frac{i}{2}}\|X\|_{\mathsf{op}} + \frac{1}{63}\frac{5}{4}d_w^{\frac{i}{2}}\sigma_{\max}(X) \\
&= \frac{80}{63}d_w^{\frac{i}{2}}\|X\|_{\mathsf{op}} \\
&\leq \frac{9}{7}d_w^{\frac{i}{2}}\|X\|_{\mathsf{op}}.
\end{aligned}
$$

This proves the last bound from the event $\mathcal{B}(t)$. $\square$

Proof of $\mathcal{A}(t)$. We show in the following that the events $\mathcal{B}(t), \mathcal{A}(t)$ imply $\mathcal{A}(t+1)$. Let us start by stating the Lemma.

Lemma A.26. *For any $\tau \leq t \leq T$, we have that $\{\{\mathcal{A}(j)\}_{\tau\leq j\leq t-1}, \{\mathcal{B}(j)\}_{\tau\leq j\leq t}\} \implies \mathcal{A}(t)$.*

Proof. From Lemma A.15, it follows that $C_{\mathsf{prod}}^{(i)} \in [\frac{1}{4}, 1]$. From this and $\mathcal{B}(t)$, it follows that

$$
\begin{aligned}
\lambda_{\min}(P(t)) &\geq \frac{1}{4d_w^{L-1}m}\sum_{i=1}^{L}\sigma_{\min}^2(W_{L:(i+1)}(t))\sigma_{\min}^2(W_{i-1:1}(t)Y) \\
&\geq \frac{1}{4d_w^{L-1}m}\sigma_{\min}^2(W_{L:2}(t))\sigma_{\min}^2(Y) \\
&\geq \frac{1}{4d_w^{L-1}m}d_w^{L-1}\left(\frac{5}{7}\right)^2(1-\delta)^2\sigma_{\min}^2(X) \\
&\geq \frac{1}{4d_w^{L-1}m}\frac{1}{2}d_w^{L-1}\frac{8}{10}\sigma_{\min}^2(X) \\
&\geq \frac{\sigma_{\min}^2(X)}{10m},
\end{aligned}
$$

given $\delta = \frac{1}{10}$. Similarly, for the upper bound on $\lambda_{\max}(P(t))$, we get

$$
\begin{aligned}
\lambda_{\max}(P(t)) &\leq \frac{1}{d_w^{L-1}m}\sum_{i=1}^{L}\sigma_{\max}^2(W_{L:(i+1)}(t))\sigma_{\max}^2(W_{i-1:1}(t)Y) \\
&\leq \frac{1}{d_w^{L-1}m}\sum_{i=1}^{L}\left(\frac{9}{7}\right)^2 d_w^{L-i}\left(\frac{9}{7}\right)^2 d_w^{i-1}\sigma_{\max}^2(X) \\
&\leq \frac{2L\sigma_{\max}^2(X)}{m}.
\end{aligned}
$$

Similarly to the first τ iterations, we obtain a bound on the higher-order terms:

$$
\begin{aligned}
&\|E(t)Y\|_{\mathsf{F}} \\
&= \|d_w^{-\frac{L-1}{2}}m^{-\frac{1}{2}}E_0(t)Y\|_{\mathsf{F}}
\end{aligned}
$$

$$
\begin{aligned}
&\le d_w^{-\frac{L-1}{2}} m^{-\frac{1}{2}} \sum_{\ell=2}^{L} \eta^\ell \binom{L}{\ell} \left(1 - \frac{\eta\lambda}{d_w}\right)^{L-\ell} \left(2\sqrt{\frac{L}{c}}\right)^{\ell-1} d_w^{\frac{L-\ell}{2}} \left(\frac{3}{\sqrt{m}} \|\Phi(t)\|_{\mathsf{F}} \|X\|_{\mathsf{op}}\right)^\ell \|Y\|_{\mathsf{op}} \\
&\le \frac{C\eta L^{\frac{3}{2}} \|X\|_{\mathsf{op}}^2 \|\Phi(t)\|_{\mathsf{F}}}{m} \sum_{\ell=1}^{L-1} \left(\frac{C\eta L^{\frac{3}{2}} \|X\|_{\mathsf{op}} \|\Phi(t)\|_{\mathsf{F}}}{(md_w)^{1/2}}\right)^\ell \\
&\lesssim \frac{\eta^2 L^3 \|X\|_{\mathsf{op}}^3 \|\Phi(t)\|_{\mathsf{F}}^2}{m^{3/2} d_w^{1/2}} \\
&\lesssim \frac{\eta^2 L^3 \lambda \|X\|_{\mathsf{op}}^4 \|\Phi(t)\|_{\mathsf{F}}}{m^2 \sigma_{\min}^2(X)} \sqrt{\frac{d}{d_w}} \\
&\le \frac{\eta L^2 \|X\|_{\mathsf{op}}^2 \|\Phi(t)\|_{\mathsf{F}}}{m} \sqrt{\frac{m}{d_w}} \\
&\le \frac{3\eta\sigma_{\min}^2(X)}{80m} \cdot \|\Phi(t)\|_{\mathsf{F}},
\end{aligned}
$$

where the second inequality follows by imitating the argument in Lemma A.14, the third inequality follows from Lemma B.5 and the inequality

$$
\begin{aligned}
\frac{C\eta L^{3/2} \|X\|_{\mathsf{op}} \|\Phi(t)\|_{\mathsf{F}}}{\sqrt{md_w}} &\overset{(\eta \le m/L\sigma_{\max}^2(X))}{\le} \frac{CL^{1/2}\sqrt{m}\|\Phi(t)\|_{\mathsf{F}}}{\sqrt{d_w}\sigma_{\max}(X)} \\
&\overset{(\mathcal{A}(t))}{\le} \frac{C\lambda L^{1/2}\sqrt{\mathrm{sr}(X)}}{\sigma_{\min}^2(X)} \sqrt{\frac{d}{d_w}} \\
&\overset{(\lambda \lesssim L\sigma_{\min}^2(X))}{\lesssim} CL^{3/2} \sqrt{\frac{d\,\mathrm{sr}(X)}{d_w}} \\
&\overset{(d_w \gtrsim L^3 d\,\mathrm{sr}(X))}{\le} \frac{1}{2},
\end{aligned}
$$

the second to last inequality uses that $\eta \le \frac{m}{L\sigma_{\max}^2(X)}$ and that $\lambda \le \sigma_{\min}^2(X)\sqrt{\frac{m}{d}}$ and the last inequality follows from $d_w \gtrsim mL^4\kappa^4$. Therefore, we arrive at the following bound on the regression error:

$$
\begin{aligned}
\|\mathrm{vec}(\Phi(t+1))\|_{\mathsf{F}} &= \|(I - \eta P(t))\|_{\mathsf{op}} \|\Phi(t)\|_{\mathsf{F}} + \|E(t)\|_{\mathsf{F}} + |C_{\mathsf{prod}} - 1| \|U(t)\|_{\mathsf{F}} \\
&\le \left(1 - \frac{\eta\sigma_{\min}^2(X)}{10m} + \frac{3\eta\sigma_{\min}^2(X)}{80m}\right) \|\Phi(t)\|_{\mathsf{F}} + |C_{\mathsf{prod}} - 1| \|U(t)\|_{\mathsf{F}} \\
&\le \left(1 - \frac{\eta\sigma_{\min}^2(X)}{16m}\right) \|\Phi(t)\|_{\mathsf{F}} + \left[\frac{(L-1)\eta\lambda}{d_w} + \frac{\eta\lambda}{m}\right] \left(\frac{5}{4}\sqrt{\frac{d}{m}}\right) \|X\|_{\mathsf{F}} \\
&\le \left(1 - \frac{\eta\sigma_{\min}^2(X)}{16m}\right) \|\Phi(t)\|_{\mathsf{F}} + \frac{5\sqrt{d}\eta\lambda}{2m\sqrt{m}} \|X\|_{\mathsf{F}}.
\end{aligned}
$$

We can split the remaining analysis into two cases:

1. If $\frac{40\lambda\|X\|_{\mathsf{F}}}{\sigma_{\min}^2(X)}\sqrt{\frac{d}{m}} \le \|\Phi(t)\|_{\mathsf{F}} \le \frac{80\lambda\|X\|_{\mathsf{F}}}{\sigma_{\min}^2(X)}\sqrt{\frac{d}{m}}$, then

$$
\|\Phi(t+1)\|_{\mathsf{F}} \le \left(1 - \frac{\eta\sigma_{\min}^2(X)}{16m}\right) \|\Phi(t)\|_{\mathsf{F}} + \frac{5\sqrt{d}\eta\lambda}{2m\sqrt{m}} \|X\|_{\mathsf{F}}
$$

$$
\begin{aligned}
&\le \|\Phi(t)\|_{\mathsf{F}} - \left(\frac{\eta\sigma_{\min}^2(X)}{16m}\right)\frac{40\sqrt{d}\lambda\|X\|_{\mathsf{F}}}{\sigma_{\min}^2(X)\sqrt{m}} + \frac{5\sqrt{d}\eta\lambda}{2m\sqrt{m}}\|X\|_{\mathsf{op}} \\
&\le \|\Phi(t)\|_{\mathsf{F}} - \left(\frac{40\eta\lambda\|X\|_{\mathsf{F}}\sqrt{d}}{16m\sqrt{m}}\right) + \frac{5\sqrt{d}\eta\lambda}{2m\sqrt{m}}\|X\|_{\mathsf{F}} \\
&= \|\Phi(t)\|_{\mathsf{F}} - \frac{\eta\lambda\|X\|_{\mathsf{F}}\sqrt{d}}{m\sqrt{m}}\left[\frac{5}{2} - \frac{5}{2}\right] \\
&\le \|\Phi(t)\|_{\mathsf{F}}.
\end{aligned}
$$

2. On the other hand, if $\|\Phi(t)\|_{\mathsf{F}} \le \frac{40\lambda\|X\|_{\mathsf{F}}}{\sigma_{\min}^2(X)}\sqrt{\frac{d}{m}}$, then

$$
\begin{aligned}
\|\Phi(t+1)\|_{\mathsf{F}} &\le \left(1 - \frac{\eta\sigma_{\min}^2(X)}{16m}\right)\|\Phi(t)\|_{\mathsf{F}} + \frac{5\sqrt{d}\eta\lambda}{2m\sqrt{m}}\|X\|_{\mathsf{F}} \\
&\le \left(1 - \frac{\eta\sigma_{\min}^2(X)}{16m}\right)\frac{40\lambda\|X\|_{\mathsf{F}}}{\sigma_{\min}^2(X)}\sqrt{\frac{d}{m}} + \frac{5\sqrt{d}\eta\lambda}{2m\sqrt{m}}\|X\|_{\mathsf{F}} \\
&\le \frac{40\lambda\|X\|_{\mathsf{F}}}{\sigma_{\min}^2(X)}\sqrt{\frac{d}{m}} + \frac{5\sqrt{d}\eta\lambda}{2m\sqrt{m}}\|X\|_{\mathsf{F}} \\
&\le \frac{41\lambda\|X\|_{\mathsf{F}}}{\sigma_{\min}^2(X)}\sqrt{\frac{d}{m}} \\
&\le \frac{80\lambda\|X\|_{\mathsf{F}}}{\sigma_{\min}^2(X)}\sqrt{\frac{d}{m}},
\end{aligned}
$$

   where the penultimate inequality follows from the requirement $\eta \le \frac{2m}{5\sigma_{\min}^2(X)}$. In particular, since we assumed $\mathcal{A}(t)$ holds, which means $\|\Phi(t)\|_{\mathsf{F}} \le \frac{80\lambda\|X\|_{\mathsf{F}}}{\sigma_{\min}^2(X)}\sqrt{\frac{d}{m}}$ then this also holds for $\|\Phi(t+1)\|_{\mathsf{F}}$.

This shows that the event $\mathcal{A}(t+1)$ holds. □

Proof of Theorem A.23. Taking the above Lemmas together, we have shown that the base case for all three events $\mathcal{A}, \mathcal{B}, \mathcal{C}$ holds. Further we have shown by induction that $\{\mathcal{A}(j), \mathcal{B}(j)\}_{\tau \le j < t} \implies \mathcal{C}(t+1)$, $\mathcal{C}(t) \implies \mathcal{B}(t)$ and $\mathcal{A}(t), \mathcal{B}(t) \implies \mathcal{A}(t+1)$. From this, the theorem follows. □

A.7 Step 3: Convergence off the subspace

In this section, we show that the off-subspace error depends on the hidden width. The on-subspace components of the weights act on the image of the subspace; the off-subspace components are in the orthogonal complement of the on-subspace components. More formally, the projection onto the subspace is defined as $P^{\perp}_{\mathrm{range}(Y)} := YY^{\ddagger}$. To determine the behavior off the subspace, we must consider the projection onto $\mathrm{range}(Y)^{\perp}$, which we denote $P^{\perp}_{\mathrm{range}(Y)}$. Note that

$$
\begin{aligned}
W_1(t+1)P^{\perp}_{\mathrm{range}(Y)} &= W_1(t)\left(1 - \frac{\eta\lambda}{m}\right)P^{\perp}_{\mathrm{range}(Y)} - \eta \cdot \frac{1}{\sqrt{d_w^{L-1}m}}W_{L:2}^{\mathsf{T}}\Phi(t)Y^{\mathsf{T}}P^{\perp}_{\mathrm{range}(Y)} \\
&= \left(1 - \frac{\eta\lambda}{m}\right)W_1(t)P^{\perp}_{\mathrm{range}(Y)}
\end{aligned}
$$

$$= \left(1 - \frac{\eta\lambda}{m}\right)^{t+1} W_1(0) P^{\perp}_{\mathrm{range}(Y)}$$

using $Y^{\mathsf{T}} P^{\perp}_{\mathrm{range}(Y)} = Y^{\mathsf{T}} P_{\mathrm{range}(Y)^{\perp}} = Y^{\mathsf{T}} P_{\ker(Y^{\mathsf{T}})} = 0$. By event $\mathcal{B}(t)$ from Equation (69b), we have

$$\begin{aligned} \|W_{L:1}(t) P^{\perp}_{\mathrm{range}(Y)}\|_{\mathsf{op}} &\leq \|W_{L:2}(t)\|_{\mathsf{op}} \|W_1(t) P^{\perp}_{\mathrm{range}(Y)}\|_{\mathsf{op}} \\ &\leq \frac{9}{7} d_w^{\frac{L-1}{2}} \cdot \left(1 - \frac{\eta\lambda}{m}\right)^{t} \|W_1(0) P^{\perp}_{\mathrm{range}(Y)}\|_{\mathsf{op}} \end{aligned} \tag{73}$$

Normalizing on both sides we obtain

$$\left\| \frac{1}{\sqrt{d_w^{L-1} m}} W_{L:1}(t) P^{\perp}_{\mathrm{range}(Y)} \right\|_{\mathsf{op}} \leq 2 \left(1 - \frac{\eta\lambda}{m}\right)^{t} \frac{1}{\sqrt{m}} \|W_1(0) P^{\perp}_{\mathrm{range}(Y)}\|_{\mathsf{op}}. \tag{74}$$

We now turn to bounding $\|W_1(0) P^{\perp}_{\mathrm{range}(Y)}\|_{\mathsf{op}}$. Let $V_{\perp} \in O(m, m-s)$ be a matrix whose columns span $\mathrm{range}(Y)^{\perp}$; by orthogonal invariance of the operator norm and the Gaussian distribution, we have

$$\begin{aligned} \|W_1(0) P^{\perp}_{\mathrm{range}(Y)}\|_{\mathsf{op}} &= \|W_1(0) V_{\perp} V_{\perp}^{\mathsf{T}}\|_{\mathsf{op}} \\ &= \|W_1(0) V_{\perp}\|_{\mathsf{op}}, \end{aligned}$$

where $W_1(0) V_{\perp} \in \mathbb{R}^{d_w \times (m-s)}$ is a matrix with standard Gaussian elements; indeed,

$$W_1(0) V_{\perp} = \begin{bmatrix} W_1(0)(V_{\perp})_{:,1} & \dots & W_1(0)(V_{\perp})_{:,m-s} \end{bmatrix} \overset{(d)}{=} \begin{bmatrix} \bar{g}_1 & \dots \bar{g}_{m-s} \end{bmatrix}, \quad \text{where } \bar{g}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_{d_w}).$$

Therefore by [59, Corollary 7.3.3], the following holds with probability $1 - 2\exp(-c d_w^2)$:

$$\|W_1(0) P^{\perp}_{\mathrm{range}(Y)}\|_{\mathsf{op}} \leq 2\sqrt{d_w} + \sqrt{m-s} \lesssim \sqrt{d_w}.$$

By the preceding displays,

$$\begin{aligned} \left\| \frac{1}{\sqrt{d_w^{L-1} m}} W_{L:1}(T) P^{\perp}_{\mathrm{range}(Y)} \right\|_{\mathsf{op}} &\lesssim \left(1 - \frac{\eta\lambda}{m}\right)^{T} \cdot \sqrt{\frac{d_w}{m}} \\ &= \left(1 - \frac{\eta\lambda}{m}\right)^{T} \exp\left(\frac{1}{2}\log(d_w/m)\right) \\ &\leq \exp\left(-\frac{T\eta\lambda}{m} + \frac{1}{2}\log(d_w)\right) \\ &= d_w^{-\frac{3}{2}}, \end{aligned} \tag{75}$$

where the second inequality follows from the identity $1 - x \leq \exp(-x)$, the penultimate inequality follows from the choice of $T = \frac{2\log(d_w)\sqrt{dm}}{\eta\gamma\sigma^2_{\min}(X)}$ from Equation (30b) and the choice of $\lambda = \gamma\sigma^2_{\min}(X)\sqrt{\frac{m}{d}}$.

A.8 Robustness at test time

Suppose we have trained our model for T steps and $W_{L:1}(T) = W_L(T) \cdots W_1(T)$ are the weights of the model at the end of training. In what follows, we suppress the iteration index T for simplicity. By Theorem 2.3,

$$\|W_{L:1}Y - X\|_{\mathsf{F}} \leq C_1\gamma\|X\|_{\mathsf{F}}; \tag{76a}$$

$$\|W_{L:1}P^{\perp}_{\mathrm{range}(Y)}\|_{\mathsf{op}} \leq d_w^{-C_2} \tag{76b}$$

for universal constants $C_1, C_2 > 0$. Suppose that we receive a new test pair (x, y) satisfying

$$y = Ax + \epsilon, \quad x \in \mathrm{range}(R), \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I_m). \tag{77}$$

The next corollary characterizes the estimation error $\|W_{L:1}y - x\|$.

Corollary A.27. *Let $(W_1, \ldots, W_L)$ be the weight matrices of a deep linear network trained for T iterations in the setting of Theorem 2.3. Consider a new data point (x, y) satisfying (77). Then the output of the network, $W_{L:1}y$, satisfies*

$$\|W_{L:1}y - x\| \lesssim \gamma\kappa\sqrt{\mathrm{sr}(X)} + \frac{1}{d_w^{C_2}} + \sigma\sqrt{s} \tag{78}$$

with probability of at least $1 - c_1\exp(-c_2 d) - \exp(-c_3 s)$. Conversely, let $(W_1^{\lambda=0}(t), ..., W_L^{\lambda=0}(t))$ be the weight matrices of a deep linear network trained in the setting of Theorem 2.3 with $\lambda = 0$. Then for any $\beta > 0$, there exists an iteration T such that the reconstruction error $\|W_{L:1}^{\lambda=0}(t)Y - X\|_{\mathsf{F}} \leq \beta\|X\|_{\mathsf{F}}$ for all $t > T$. Moreover, with probability at least $1 - c_1\exp(-c_2 d) - \exp(-c_3 s)$, the test error satisfies

$$\|W_{L:1}^{\lambda=0}(t)y - x\| \gtrsim \sigma\left(\sqrt{\frac{d(m-s)}{m}} - \sqrt{s}\right) - \beta\kappa\sqrt{\mathrm{sr}(X)}\|y\|.$$

Proof. We start by bounding the error between the "oracle" solution mapping $XY^{\dagger}$ and the trained neural network. We have

$$\begin{aligned}
\|W_{L:1} - XY^{\dagger}\|_{\mathsf{op}} &= \|W_{L:1}YY^{\dagger} - XY^{\dagger} + W_{L:1}(I - YY^{\dagger})\|_{\mathsf{op}} \\
&\leq \|W_{L:1}Y - X\|_{\mathsf{op}}\|Y^{\dagger}\|_{\mathsf{op}} + \|W_{L:1}(I - YY^{\dagger})\|_{\mathsf{op}} \\
&\leq \|W_{L:1}Y - X\|_{\mathsf{F}}\|Y^{\dagger}\|_{\mathsf{op}} + \|W_{L:1}P^{\perp}_{\mathrm{range}(Y)}\|_{\mathsf{op}} \\
&\leq C_1\gamma\|X\|_{\mathsf{F}}\|Y^{\dagger}\|_{\mathsf{op}} + d_w^{-C_2} \\
&= \frac{C_1\gamma\sqrt{\mathrm{sr}(X)}\|X\|_{\mathsf{op}}}{\sigma_{\min}(Y)} + d_w^{-C_2} \\
&\leq \frac{C_1\gamma\sqrt{\mathrm{sr}(X)}\sigma_{\max}(X)}{(1-\delta)\sigma_{\min}(X)} + d_w^{-C_2} \\
&\lesssim \gamma\kappa\sqrt{\mathrm{sr}(X)} + d_w^{-C_2},
\end{aligned}$$

where the third inequality follows from Equations (76a) and (76b), the second equality follows from the definition of $\mathrm{sr}(X)$ and the identity $\|Y^{\dagger}\|_{\mathsf{op}} = \frac{1}{\sigma_{\min}(Y)}$, the penultimate inequality follows from Assumption 2.1 and the last inequality follows by substituting $\delta = \frac{1}{10}$. Consequently, we have

$$\begin{aligned}
\|W_{L:1}y - x\| &= \|(W_{L:1} - XY^{\dagger})y + XY^{\dagger}y - x\| \\
&\leq \|W_{L:1} - XY^{\dagger}\|_{\mathsf{op}}\|y\| + \|XY^{\dagger}y - x\| \\
&\lesssim \left(\gamma\kappa\sqrt{\mathrm{sr}(X)} + d_w^{-C_2}\right)\|y\| + \|XY^{\dagger}y - x\|. \qquad (79)
\end{aligned}$$

We now argue that the second term in (79) is bounded by $\sigma\sqrt{s}$. Recall that $Y = AX$, $X = RZ$ for some $Z \in \mathbb{R}^{s\times n}$ with full row rank, and $x = Rz$ for some $z \in \mathbb{R}^s$. Therefore, we have

$$\begin{aligned}
XY^{\dagger}y &= RZ(ARZ)^{\dagger}y \\
&= RZZ^{\dagger}(AR)^{\dagger}y \\
&= R(AR)^{\dagger}(AR)z + R(AR)^{\dagger}\epsilon \\
&= Rz + R(AR)^{\dagger}\epsilon \\
&= x + R(AR)^{\dagger}\epsilon.
\end{aligned}$$

The second equality in the preceding display follows from the fact that $(M_1M_2)^{\dagger} = M_2^{\dagger}M_1^{\dagger}$ when M_1 and M_2 are full column-rank and full row-rank respectively; indeed, here $M_1 \equiv AR$ is full column-rank by Assumption 2.1 and $M_2 \equiv Z$ is full row-rank by assumption. Similarly, the third and fourth inequalities follow from the full row rankness and full column rankness of Z and AR, respectively. Consequently, we have the bound

$$\|XY^{\dagger}y - x\| = \|R(AR)^{\dagger}\epsilon\| = \|(AR)^{\dagger}\epsilon\|,$$

since R is a matrix with orthogonal columns. We now write

$$AR = \bar{U}\bar{\Sigma}\bar{V}^{\mathsf{T}}, \quad \text{where } \bar{U} \in O(m, s),\ \bar{V} \in O(s),\ 1 - \delta \leq \Sigma_{ii} \leq 1 + \delta,$$

for the economic SVD of AR, where the inequalities on the singular values follow from Assumption 2.1. In particular,

$$\|(AR)^{\dagger}\epsilon\| = \|\bar{V}\bar{\Sigma}^{-1}\bar{U}^{\mathsf{T}}\epsilon\| \leq \frac{1}{\sigma_{\min}(\bar{\Sigma})}\|\bar{U}^{\mathsf{T}}\epsilon\| \lesssim \|\bar{U}^{\mathsf{T}}\epsilon\|,$$

using $\delta = \frac{1}{10}$ in the last inequality. Finally, by standard properties of the multivariate normal distribution,

$$\bar{U}^{\mathsf{T}}\epsilon \sim \mathcal{N}(0, \sigma^2 I_s) \implies \|\bar{U}^{\mathsf{T}}\epsilon\| \lesssim \sigma\sqrt{s},$$

with probability at least $1 - \exp(-cs^2)$ [59, Theorem 3.1.1], for a universal constant $c > 0$. Returning to (79), we conclude that

$$\|W_{L:1}y - x\| \lesssim \left(\gamma\kappa\sqrt{\mathrm{sr}(X)} + d_w^{-C_2}\right)\|y\| + \sigma\sqrt{s},$$

with probability at least $1 - c_1\exp(-c_2d) - \exp(-c_3s)$. This proves the first of the two claims.

We now prove the lower bound for the reconstruction error for the weights $W_i^{\lambda=0}(t)$. For simplicity, we write $\bar{W}_{L:1} := W_L^{\lambda=0}(t)\dots W_1^{\lambda=0}(t)$ and suppress the dependence on t. We obtain

$$\|\bar{W}_{L:1}y - XY^{\dagger}y\| = \|\bar{W}_{L:1}(I - YY^{\dagger})y + (\bar{W}_{L:1}Y - X)Y^{\dagger}y\|$$

$$\begin{aligned}
&\geq \|\bar{W}_{L:1}P^{\perp}_{\mathrm{range}(Y)}y\| - \frac{\|(\bar{W}_{L:1}Y - X)\|_{\mathsf{F}}}{\sigma_{\min}(Y)}\|y\| \\
&\geq \|\bar{W}_{L:1}P^{\perp}_{\mathrm{range}(Y)}\epsilon\| - \frac{\beta\|X\|_{\mathsf{F}}}{\sigma_{\min}(Y)}\|y\| \\
&\gtrsim \sqrt{\frac{d}{m}}\|P^{\perp}_{\mathrm{range}(Y)}\epsilon\| - \frac{\beta\sqrt{\mathrm{sr}(X)}\sigma_{\max}(X)}{(1-\delta)\sigma_{\min}(X)}\|y\| \\
&\gtrsim \sigma\sqrt{\frac{d(m-s)}{m}} - \beta\kappa\sqrt{\mathrm{sr}(X)}\|y\|,
\end{aligned} \tag{80}$$

where the first inequality follows from the reverse triangle inequality and the identity $\|Y^{\dagger}\| = {}^{1}/_{\sigma_{\min}(Y)}$, the second inequality follows by the assumption that $t > T$, the third inequality follows from Assumption 2.1, the definition of $\mathrm{sr}(X)$ and Lemma B.2 combined with property $\mathcal{C}(t)$ from Appendix A.5, and the last inequality follows from the fact that

$$\|P^{\perp}_{\mathrm{range}(Y)}\epsilon\| \gtrsim \sigma\sqrt{m-s}, \quad \text{with probability at least } 1 - \exp(-c(m-s)^2). \tag{81}$$

To see (81), let $V_{\perp} \in O(m, m-s)$ be a matrix whose columns span $\mathrm{range}(Y)^{\perp}$ such that $P^{\perp}_{\mathrm{range}(Y)} = V_{\perp}V_{\perp}^{\mathsf{T}}$. By orthogonal invariance of the Gaussian distribution,

$$V_{\perp}^{\mathsf{T}}\epsilon \overset{(d)}{=} \mathcal{N}(0, \sigma^2 I_{m-s}).$$

Moreover, by orthogonal invariance of the Euclidean norm,

$$\|P^{\perp}_{\mathrm{range}(Y)}\epsilon\| = \|V_{\perp}^{\mathsf{T}}\epsilon\|.$$

Combining the two preceding displays with [59, Theorem 3.1.1] yields the inequality (81).

Altogether, we get the following lower bound

$$\begin{aligned}
\|\bar{W}_{L:1}(y) - x\| &= \|\bar{W}_{L:1}y - XY^{\dagger}y + XY^{\dagger}y - x\| \\
&\geq \|(\bar{W}_{L:1} - XY^{\dagger})y\| - \|XY^{\dagger}y - x\| \\
&\gtrsim \sigma\sqrt{\frac{d(m-s)}{m}} - \beta\kappa\sqrt{s}\|y\| - \|XY^{\dagger}Ax - x\| - \|XY^{\dagger}\epsilon\| \\
&\gtrsim \sigma\sqrt{\frac{d(m-s)}{m}} - \beta\kappa\sqrt{\mathrm{sr}(X)}\|y\| - \sigma\sqrt{s} \\
&= \sigma\left(\sqrt{\frac{d(m-s)}{m}} - \sqrt{s}\right) - \beta\kappa\sqrt{\mathrm{sr}(X)}\|y\|,
\end{aligned}$$

where the first inequality follows from the reverse triangle inequality, the second inequality follows from the bound (80) and the last inequality follows from the fact that $XY^{\dagger}Ax = x$ and the upper bound $\|XY^{\dagger}\epsilon\| \lesssim \sigma\sqrt{s}$, which follows from standard properties of the multivariate Gaussian distribution. This lower bound holds with probability at least $1 - c_1\exp(-c_2 d) - \exp(-c_3 s)$. This concludes the proof of the lower bound. □

B Auxiliary results

In this section, we state and prove results used to prove the main result Theorem A.2 or mentioned in the introduction. We start with a result showing that a global minimizer solution of the regularized optimization problem is zero on the orthogonal complement of the image.

Lemma B.1. *Suppose $f_{W_{L:1}}$ is a global minimizer of the regularized optimization problem* (2). *Then $f_{W_{L:1}}$ satisfies $W_1 P^{\perp}_{\mathrm{range}(Y)} = 0$, where $P^{\perp}_{\mathrm{range}(Y)}$ is the projection onto the orthogonal complement of* $\mathrm{range}(Y)$*.*

Proof. Suppose that $f_{W_{L:1}}$ is a minimizer with $W_1 P^{\perp}_{\mathrm{range}(Y)} \neq 0$. Then consider $f_{W_{L:1} P_{\mathrm{range}(Y)}}$, the neural network that coincides with $f_{W_{L:1}}$ except that its first-layer weights are right-multiplied by $P_{\mathrm{range}(Y)}$. We have

$$\|f_{W_{L:1} P_{\mathrm{range}(Y)}}(Y) - X\|_{\mathsf{F}} = \|f_{W_{L:1}}(P_{\mathrm{range}(Y)} Y) - X\|_{\mathsf{F}} = \|f_{W_{L:1}}(Y) - X\|_{\mathsf{F}}.$$

Hence the first term in the objective in (2) is the same for $f_{W_{L:1}}$ and $f_{W_{L:1} P_{\mathrm{range}(Y)}}$. By the Pythagorean theorem, we have that

$$\|W_1\|_{\mathsf{F}}^2 = \|W_1 P_{\mathrm{range}(Y)}\|_{\mathsf{F}}^2 + \|W_1 P^{\perp}_{\mathrm{range}(Y)}\|_{\mathsf{F}}^2 > \|W_1 P_{\mathrm{range}(Y)}\|_{\mathsf{F}}^2$$

since $W_1 P^{\perp}_{\mathrm{range}(Y)} \neq 0$ by assumption. Thus the regularization term in the objective (2) is strictly larger for $f_{W_{L:1}}$ than for $f_{W_{L:1} P_{\mathrm{range}(Y)}}$. Therefore $f_{W_{L:1}}$ cannot be the minimal-norm solution. $\square$

Lemma B.2. *Let $A_1, A_2, \ldots, A_q$ have i.i.d. Gaussian elements with $A_i \in \mathbb{R}^{n_i \times n_{i-1}}$, $n_0 = n$, and $n_i \gtrsim q$. Then*

$$\mathbb{E}\left[\|A_q \ldots A_1 y\|^2\right] = \|y\|^2 \cdot \prod_{i=1}^{q} n_i, \tag{82}$$

$$\mathbb{P}\left\{\left|\|A_q \cdots A_1 y\|^2 - \|y\|^2 \prod_{i=1}^{q} n_i\right| \geq 0.1\|y\|^2 \prod_{i=1}^{q} n_i\right\} \leq c_1 \exp\left(-\frac{c_2}{\sum_{i=1}^{q} n_i^{-1}}\right), \tag{83}$$

where $c_1, c_2 > 0$ are universal constants and y is any fixed vector.

Proof. We start with (82). Note that for any A_i, we have

$$\begin{aligned}
\|A_i y\|^2 &= \sum_{j=1}^{n_i} \langle (A_i)_{j,:}, y\rangle^2 \\
&\overset{(d)}{=} \sum_{j=1}^{n_i} \|y\|^2 g_i^2 \qquad (g_i \sim \mathcal{N}(0,1)) \\
&\overset{(d)}{=} \|y\|^2 Z_i,
\end{aligned}$$

where $Z_i \sim \chi^2_{n_i}$, a χ^2-random variable with n_i degrees of freedom. As a result,

$$\mathbb{E}\left[\|A_i y\|^2\right] = \|y\|^2 \mathbb{E}\left[Z_i\right] = \|y\|^2 \cdot n_i.$$

Moreover, since $A_1, \ldots, A_q$ are independent, we have

$$\begin{aligned}
\mathbb{E}\left[\|A_q \ldots A_1 y\|^2\right] &= \mathbb{E}\left[\mathbb{E}\left[\|A_q (A_{q-1} \ldots A_1 y)\|^2 \mid A_1, \ldots, A_{q-1}\right]\right] \\
&= n_q \mathbb{E}\left[\|A_{q-1} \ldots A_1 y\|^2\right] \\
&= n_q \mathbb{E}\left[\mathbb{E}\left[\|A_{q-1} \ldots A_1 y\|^2 \mid A_1, \ldots, A_{q-2}\right]\right] \\
&= n_q \cdot n_{q-1} \mathbb{E}\left[\|A_{q-2} \ldots A_1 y\|^2\right]
\end{aligned}$$

$$
\begin{aligned}
&= \dots \\
&= \prod_{j=1}^{q} n_i \cdot \|y\|^2,
\end{aligned}
$$

by iterating the above construction; this proves Equation (82).

We now prove Equation (83). Let $\|y\| = 1$ for simplicity; then $\|A_q \dots A_1 y\|^2 \sim Z_q Z_{q-1} \dots Z_1$, where $Z_i \sim \chi^2_{n_i}$. The moments of a random variable $X \sim \chi^2_k$ satisfy

$$
\begin{aligned}
\mathbb{E}[X^\lambda] &= \frac{2^\lambda \Gamma(\frac{k}{2} + \lambda)}{\Gamma(\frac{k}{2})} \\
&= \frac{2^\lambda \sqrt{\frac{4\pi}{k+2\lambda}} \left(\frac{k+2\lambda}{2e}\right)^{\frac{k}{2}+\lambda}}{\sqrt{\frac{4\pi}{k}} \left(\frac{k}{2e}\right)^{\frac{k}{2}}} \left(1 + O(1/k)\right),
\end{aligned}
$$

for all $\lambda > -k/2$, with the second equality furnished by a Stirling approximation. Following [Eq. (20)] in [36], we obtain the following upper bound:

$$
\mathbb{E}[X^\lambda] \leq \exp\left(\frac{2\lambda^2}{k} - \frac{1}{2}\log\left(1 + \frac{2\lambda}{k}\right) + \lambda \log k\right) \cdot \left(1 + O\left(\frac{1}{k}\right)\right), \quad \forall \lambda \geq -\frac{k}{4}. \tag{84}
$$

To bound the upper tail in Equation (83), we argue that for any $\lambda > 0$,

$$
\begin{aligned}
&\mathbb{P}\left\{Z_q \dots Z_1 \geq \exp(c) \prod_{i=1}^{q} n_i\right\} \\
&\leq \exp(-\lambda c) \left(\prod_{i=1}^{q} n_i\right)^{-\lambda} \cdot \mathbb{E}[(Z_q \dots Z_1)^\lambda] \\
&= \exp\left(-\lambda c - \lambda \log\left(\prod_{i=1}^{q} n_i\right)\right) \mathbb{E}[(Z_q \dots Z_1)^\lambda] \\
&\leq \exp\left(-\lambda c - \lambda \log\left(\prod_{i=1}^{q} n_i\right) + \sum_{i=1}^{q} \frac{2\lambda^2}{n_i} + \lambda \log(n_i) - \frac{1}{2}\log\left(1 + \frac{2\lambda}{n_i}\right)\right) \prod_{j=1}^{q} \left(1 + O\left(\frac{1}{n_i}\right)\right).
\end{aligned} \tag{85}
$$

Under our assumption that $n_i \gtrsim q$, the last term above satisfies

$$
\begin{aligned}
\prod_{j=1}^{q} \left(1 + O\left(\frac{1}{n_i}\right)\right) &\lesssim \prod_{j=1}^{q} \left(1 + \frac{1}{q}\right) \\
&= \left(1 + \frac{1}{q}\right)^q \\
&\leq \lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n \\
&= \exp(1),
\end{aligned} \tag{86}
$$

using the formal definition of the exponential. For the upper tail, the exponent in (85) can be simplified as follows:

$$-\lambda c + \sum_{i=1}^{q} \frac{2\lambda^2}{n_i} - \frac{1}{2}\log\left(1 + \frac{2\lambda}{n_i}\right) \le -\lambda c + 2\lambda^2 \sum_{i=1}^{q} \frac{1}{n_i},$$

using $\log(1+u) \ge 0$ for any $u \ge 0$. Maximizing over $\lambda \ge 0$ yields

$$\lambda_\star = \frac{c}{4 \cdot \sum_{i=1}^{q} \frac{1}{n_i}}.$$

Plugging the value of $\lambda_\star$ into the upper bound for the exponent leads to

$$\begin{aligned}
&-\lambda_\star c + 2\lambda_\star^2 \sum_{i=1}^{q} \frac{1}{n_i} \\
&= -\frac{c^2}{4} \frac{1}{\sum_{i=1}^{q} n_i^{-1}} + \frac{c^2}{8} \cdot \frac{\sum_{i=1}^{q} \frac{1}{n_i}}{\left(\sum_{i=1}^{q} \frac{1}{n_i}\right)^2} \\
&= -\frac{c^2}{8} \cdot \frac{1}{\sum_{i=1}^{q} n_i^{-1}}.
\end{aligned}$$

Setting $c = \log(1.1)$ completes the proof.

We now derive the lower bound in Equation (83). Given $\lambda < 0$, we have

$$\begin{aligned}
&\mathbb{P}\left\{ Z_q \dots Z_1 \le \exp(-c) \prod_{i=1}^{q} n_i \right\} \\
&= \mathbb{P}\left\{ (Z_q \dots Z_1)^\lambda \ge \exp(-\lambda c) \Big(\prod_{i=1}^{q} n_i \Big)^\lambda \right\} \\
&\le \exp\left(\lambda c - \lambda \log\Big(\prod_{i=1}^{q} n_i \Big) \right) \mathbb{E}[(Z_q \dots Z_1)^\lambda] \\
&\le C_1 \exp\left(\lambda c - \lambda \sum_{i=1}^{q} \log(n_i) + \sum_{i=1}^{q} \frac{2\lambda^2}{n_i} - \frac{1}{2}\log\left(1 + \frac{2\lambda}{n_i}\right) + \lambda \log(n_i) \right) \\
&= C_1 \exp\left(\lambda c + 2\lambda^2 \sum_{i=1}^{q} \frac{1}{n_i} - \frac{1}{2}\log\left(1 + \frac{2\lambda}{n_i}\right) \right)
\end{aligned}$$

In particular, the exponent in the preceding display satisfies

$$\lambda c + \sum_{i=1}^{q} \frac{2\lambda^2}{n_i} - \frac{1}{2}\log\left(1 + \frac{2\lambda}{n_i}\right) \le \lambda c + 2\lambda^2 \sum_{i=1}^{q} \frac{1}{n_i} - 2\lambda \sum_{i=1}^{q} \frac{1}{n_i},$$

using the inequaliy $\log(1+2x) \ge 4x$ valid for any $x > -\frac{1}{4}$. Setting $\lambda = -\frac{c}{4\sum_{i=1}^{q} n_i^{-1}}$ yields

$$-\frac{c^2}{4\sum_{i=1}^{q} n_i^{-1}} + \frac{c^2}{8} \frac{\sum_{i=1}^{q} n_i^{-1}}{(\sum_{i=1}^{q} n_i^{-1})^2} + \frac{c}{2} = -\frac{c^2}{4\sum_{i=1}^{q} n_i^{-1}} + \frac{c}{2}.$$

Setting $c = -\log(0.9)$ completes the proof. ∎

Theorem B.3 (Weierstrass)**.** *The following inequality holds:*

$$1 - \sum_{i=1}^{n} w_i x_i \leq \prod_{i=1}^{n} (1 - x_i)^{w_i}, \quad \textit{for all } x \in [0,1] \textit{ and } w_i \geq 1. \tag{87}$$

Proof. We prove the inequality by induction on the number of terms. For the base case $n = 1$, consider the function

$$h(w) = (1 - x_1)^w - (1 - w x_1), \quad \text{with } \; h'(w) = (1 - x_1)^w \log(1 - x_1) + x_1$$

Clearly $h(1) = 0$, so it suffices to show h is increasing on $[1, \infty)$. Starting from the inequality $\log(1 - x_1) \geq \frac{x_1}{x_1 - 1}$, we have

$$\begin{aligned} h'(w) &\geq \frac{x_1(1 - x_1)^w}{x_1 - 1} + x_1 \\ &= \frac{x_1(1 - x_1)^w + x_1(x_1 - 1)}{x_1 - 1} \\ &= \frac{x_1\left[(1 - x_1) - (1 - x_1)^w\right]}{1 - x_1} \\ &\geq 0, \quad \text{for all } w \geq 1, \end{aligned}$$

since $(1 - x_1) \in (0, 1)$. This proves the claim for $n = 1$.

Now suppose the claim holds up to some $n \in \mathbb{N}$. We have

$$\begin{aligned} \prod_{j=1}^{n+1} (1 - x_j)^{w_j} &= (1 - x_{n+1})^{w_{n+1}} \prod_{j=1}^{n} (1 - x_j)^{w_j} \\ &\geq (1 - w_{n+1} x_{n+1}) \prod_{j=1}^{n} (1 - x_j)^{w_j} \\ &\geq (1 - w_{n+1} x_{n+1}) \left(1 - \sum_{j=1}^{n} w_j x_j\right) \\ &= 1 - \sum_{j=1}^{n+1} w_j x_j + w_{n+1} x_{n+1} \cdot \sum_{j=1}^{n} w_j x_j \\ &\geq 1 - \sum_{j=1}^{n+1} w_j x_j, \end{aligned}$$

where the first inequality follows from the base case, the second inequality follows by the inductive hypothesis and the last inequality follows from nonnegativity of $\{w_j\}_{j \geq 1}$ and $\{x_j\}_{j \geq 1}$. This completes the proof. □

Lemma B.4. *Under the assumptions of Theorem A.2, we have that*

$$|1 - C_{\mathsf{prod}}| \leq \frac{(L-1)\eta\lambda}{d_w} + \frac{\eta\lambda}{m} \leq \frac{2\eta\lambda}{m}. \tag{88}$$

Proof. Since $C_{\mathsf{prod}} < 1$, we have $|C_{\mathsf{prod}} - 1| = 1 - \prod_{i=1}^{L} \left(1 - \frac{\eta\lambda}{d_i}\right)$. Now, let $x_i := \frac{\eta\lambda}{d_i}$ and $w_i := 1$ for $i = 1, \dots, L$. From Theorem B.3, it follows that

$$1 - \prod_{i=1}^{L} \left(1 - \frac{\eta\lambda}{d_i}\right) \le \sum_{i=1}^{L} \frac{\eta\lambda}{d_i} = \frac{(L-1)\eta\lambda}{d_w} + \frac{\eta\lambda}{m} \le \frac{2\eta\lambda}{m},$$

under the assumption that $d_w \ge m(L-1)$. □

Lemma B.5. *For any $\alpha \le \frac{1}{2}$ and $j, k \in \mathbb{N}$, it holds that*

$$\sum_{i=j}^{k} \alpha^i \le 2\alpha^j (1 - \alpha^{k-j+1}). \tag{89}$$

Proof. The claim follows from the geometric series formula:

$$\sum_{i=j}^{k} \alpha^i = \alpha^j \sum_{i=0}^{k-j} \alpha^i = \alpha^j \cdot \frac{1 - \alpha^{k-j+1}}{1 - \alpha} \le 2\alpha^j (1 - \alpha^{k-j+1}),$$

where the last inequality follows from $1/1-\alpha \le 2$. □

C Information on numerics for the union of subspaces model

Data generation for the union of subspaces experiments. The union-of-subspaces model stipulates that each vector in the input data belongs to one of k subspaces. Formally, there exists a collection $\mathcal{R} := \{R_1, \dots, R_k\}$, where $R_i \in O(d, s)$, such that $x^i \in \bigcup_{j=1}^{k} \mathrm{range}(R_j)$ for all i. In our experiments, we generate training samples from the union-of-subspaces model in the following manner:

- Sample $Z \in \mathbb{R}^{d \times n}$ according to the procedure described in Section 3.
- For each $i = 1, \dots, n$:
  1. Sample $R \sim \mathrm{Unif}(\mathcal{R})$.
  2. Set $X_{:,i} = RZ_{:,i}$.

Neural network architecture for the union-of-subspaces model. The inverse mapping for linear inverse problems with data from a union-of-subspaces model is in general nonlinear for $k > 1$ — as a result, deep linear networks are not a suitable choice for learning the inverse mapping. Nevertheless, it is known that the inverse mapping is approximated to arbitrary accuracy by a piecewise-linear mapping (see [60]), which can be realized as a multi-index model of the form $g(V^\mathsf{T} x)$ for suitable V and vector-valued mapping g. Guided by this, we use a neural network architecture defined as follows:

$$f_{W_1, \dots, W_L}(x) = W_L \left(W_{L-1} W_{L-2} \cdots W_1 x\right)_+, \tag{90}$$

where $W_1, \dots, W_L$ are learnable weight matrices and $[\cdot]_+$ denotes the (elementwise) positive part, equivalent to using a ReLU activation at the $(L-1)^{\text{th}}$ hidden layer. Indeed, recent results [15] suggest that neural networks of the form (90) are biased towards multi-index models such as the one sought to approximate the inverse mapping. Finally, all the networks from Figure 1 were trained for 100000 iterations with learning rate $\eta = 10^{-3}$.