Implicit Regularization in Deep Learning May Not Be Explainable by Norms

Noam Razin
Tel Aviv University
noam.razin@cs.tau.ac.il

Nadav Cohen
Tel Aviv University
cohennadav@cs.tau.ac.il

Abstract

Mathematically characterizing the implicit regularization induced by gradient-based optimization is a longstanding pursuit in the theory of deep learning. A widespread hope is that a characterization based on minimization of norms may apply, and a standard test-bed for studying this prospect is matrix factorization (matrix completion via linear neural networks). It is an open question whether norms can explain the implicit regularization in matrix factorization. The current paper resolves this open question in the negative, by proving that there exist natural matrix factorization problems on which the implicit regularization drives *all* norms (and quasi-norms) *towards infinity*. Our results suggest that, rather than perceiving the implicit regularization via norms, a potentially more useful interpretation is minimization of rank. We demonstrate empirically that this interpretation extends to a certain class of non-linear neural networks, and hypothesize that it may be key to explaining generalization in deep learning.[1]

1 Introduction

A central mystery in deep learning is the ability of neural networks to generalize when having far more learnable parameters than training examples. This generalization takes place even in the absence of any explicit regularization (see [88]), thus a view by which gradient-based optimization induces an *implicit regularization* has arisen (see, *e.g.*, [64]). Mathematically characterizing this implicit regularization is regarded as a major open problem in the theory of deep learning (*cf.* [66]). A widespread hope (initially articulated in [65]) is that a characterization based on *minimization of norms* (or quasi-norms[2]) may apply. Namely, it is known that for linear regression, gradient-based optimization converges to solution with minimal ℓ_2 norm (see for example Section 5 in [88]), and the hope is that this result can carry over to neural networks if we allow ℓ_2 norm to be replaced by a different (possibly architecture- and optimizer-dependent) norm (or quasi-norm).

A standard test-bed for studying implicit regularization in deep learning is *matrix completion* (*cf.* [34, 8]): given a randomly chosen subset of entries from an unknown matrix W^*, the task is to recover the unseen entries. This may be viewed as a prediction problem, where each entry in W^* stands for a data point: observed entries constitute the training set, and the average reconstruction error over the unobserved entries is the test error, quantifying generalization. Fitting the observed entries is obviously an underdetermined problem with multiple solutions. However, an extensive body of work (see [26] for a survey) has shown that if W^* is low-rank, certain technical assumptions (*e.g.* "incoherence") are satisfied and sufficiently many entries are observed, then various algorithms can achieve approximate or even exact recovery. Of these, a well-known method based upon convex optimization finds the minimal nuclear norm[a] matrix among those fitting observations (see [15]).

[1] Due to lack of space, a significant portion of the paper is deferred to the appendices. We refer the reader to [72] for a self-contained version of the text.

[2] A *quasi-norm* $\|\cdot\|$ on a vector space $\mathcal{V}$ is a function from $\mathcal{V}$ to $\mathbb{R}_{\geq 0}$ that satisfies the same axioms as a norm, except for the triangle inequality $\forall v_1, v_2 \in \mathcal{V} : \|v_1 + v_2\| \leq \|v_1\| + \|v_2\|$, which is replaced by the weaker requirement $\exists c \geq 1 \;\; s.t. \;\; \forall v_1, v_2 \in \mathcal{V} : \|v_1 + v_2\| \leq c \cdot (\|v_1\| + \|v_2\|)$.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

One may try to solve matrix completion using shallow neural networks. A natural approach, *matrix factorization*, boils down to parameterizing the solution as a product of two matrices — $W = W_2W_1$ — and optimizing the resulting (non-convex) objective for fitting observations. Formally, this can be viewed as training a depth 2 linear neural network. It is possible to explicitly constrain the rank of the produced solution by limiting the shared dimension of W_1 and W_2. However, Gunasekar *et al.* have shown in [34] that in practice, even when the rank is unconstrained, running gradient descent with small learning rate (step size) and initialization close to the origin (zero) tends to produce low-rank solutions, and thus allows accurate recovery if W^* is low-rank. Accordingly, they conjectured that the implicit regularization in matrix factorization boils down to minimization of nuclear norm:

Conjecture 1 (from [34], informally stated)**.** *With small enough learning rate and initialization close enough to the origin, gradient descent on a full-dimensional matrix factorization converges to a minimal nuclear norm solution.*

In a subsequent work — [8] — Arora *et al.* considered *deep matrix factorization*, obtained by adding depth to the setting studied in [34]. Namely, they considered solving matrix completion by training a depth L linear neural network, *i.e.* by running gradient descent on the parameterization $W = W_L W_{L-1} \cdots W_1$, with $L \in \mathbb{N}$ arbitrary (and the dimensions of $\{W_l\}_{l=1}^L$ set such that rank is unconstrained). It was empirically shown that deeper matrix factorizations (larger L) yield more accurate recovery when W^* is low-rank. Moreover, it was conjectured that the implicit regularization, for any depth $L \geq 2$, can *not* be described as minimization of a mathematical norm (or quasi-norm):

Conjecture 2 (based on [8], informally stated)**.** *Given a (shallow or deep) matrix factorization, for any norm (or quasi-norm) $\|\cdot\|$, there exists a set of observed entries with which small learning rate and initialization close to the origin can* not *ensure convergence of gradient descent to a minimal (in terms of $\|\cdot\|$) solution.*

Conjectures 1 and 2 contrast each other, and more broadly, represent opposing perspectives on the question of whether norms may be able to explain implicit regularization in deep learning. In this paper, we resolve the tension between the two conjectures by affirming the latter. In particular, we prove that there exist natural matrix completion problems where fitting observations via gradient descent on a depth $L \geq 2$ matrix factorization leads — with probability 0.5 or more over (arbitrarily small) random initialization — *all* norms (and quasi-norms) to *grow towards infinity*, while the rank essentially decreases towards its minimum. This result is in fact stronger than the one suggested by Conjecture 2, in the sense that: *(i)* not only is each norm (or quasi-norm) disqualified by some setting, but there are actually settings that jointly disqualify all norms (and quasi-norms); and *(ii)* not only are norms (and quasi-norms) not necessarily minimized, but they can grow towards infinity. We corroborate the analysis with empirical demonstrations.

Our findings imply that, rather than viewing implicit regularization in (shallow or deep) matrix factorization as minimizing a norm (or quasi-norm), a potentially more useful interpretation is *minimization of rank*. As a step towards assessing the generality of this interpretation, we empirically explore an extension of matrix factorization to *tensor factorization*.[3] Our experiments show that in analogy with matrix factorization, gradient descent on a tensor factorization tends to produce solutions with low rank, where rank is defined in the context of tensors.[4] Similarly to how matrix factorization corresponds to a linear neural network whose input-output mapping is represented by a matrix, it is known (see [22]) that tensor factorization corresponds to a *convolutional arithmetic circuit* (certain type of *non-linear* neural network) whose input-output mapping is represented by a tensor. We thus obtain a second exemplar of a neural network architecture whose implicit regularization strives to lower a notion of rank for its input-output mapping. This leads us to believe that the phenomenon may be general, and formalizing notions of rank for input-output mappings of contemporary models may be key to explaining generalization in deep learning.

The remainder of the paper is organized as follows. Section 2 presents the deep matrix factorization model. Section 3 delivers our analysis, showing that its implicit regularization can drive all norms to infinity. Experiments, with both the analyzed setting and tensor factorization, are given in Section 4. For conciseness, we defer our summary to Appendix A, and review related work in Appendix B.

[3] For the sake of this paper, *tensors* can be thought of as N-dimensional arrays, with $N \in \mathbb{N}$ arbitrary (matrices correspond to the special case $N = 2$).

[4] The *rank of a tensor* is the minimal number of summands required to express it, where each summand is an outer product between vectors.

2 Deep matrix factorization

Suppose we would like to complete a d-by-d' matrix based on a set of observations $\{b_{i,j} \in \mathbb{R}\}_{(i,j)\in\Omega}$, where $\Omega \subset \{1, 2, \ldots, d\} \times \{1, 2, \ldots, d'\}$. A standard (underdetermined) loss function for the task is:

$$\ell : \mathbb{R}^{d,d'} \to \mathbb{R}_{\geq 0} \quad , \quad \ell(W) = \frac{1}{2}\sum\nolimits_{(i,j)\in\Omega} \big((W)_{i,j} - b_{i,j}\big)^2 \, . \tag{1}$$

Employing a depth L matrix factorization, with hidden dimensions $d_1, d_2, \ldots, d_{L-1} \in \mathbb{N}$, amounts to optimizing the *overparameterized objective*:

$$\phi(W_1, W_2, \ldots, W_L) := \ell(W_{L:1}) = \frac{1}{2}\sum\nolimits_{(i,j)\in\Omega} \big((W_{L:1})_{i,j} - b_{i,j}\big)^2 \, , \tag{2}$$

where $W_l \in \mathbb{R}^{d_l, d_{l-1}}$, $l = 1, 2, \ldots, L$, with $d_L := d, d_0 := d'$, and:

$$W_{L:1} := W_L W_{L-1} \cdots W_1 \, , \tag{3}$$

referred to as the *product matrix* of the factorization. Our interest lies on the implicit regularization of gradient descent, *i.e.* on the type of product matrices (Equation (3)) it will find when applied to the overparameterized objective (Equation (2)). Accordingly, and in line with prior work (*cf.* [34, 8]), we focus on the case in which the search space is unconstrained, meaning $\min\{d_l\}_{l=0}^{L} = \min\{d_0, d_L\}$ (rank is not limited by the parameterization).

As a theoretical surrogate for gradient descent with small learning rate and near-zero initialization, similarly to [34] and [8] (as well as other works analyzing linear neural networks, *e.g.* [75, 6, 53, 7]), we study *gradient flow* (gradient descent with infinitesimally small learning rate):[b]

$$\dot{W}_l(t) := \tfrac{d}{dt} W_l(t) = -\tfrac{\partial}{\partial W_l}\phi(W_1(t), W_2(t), \ldots, W_L(t)) \quad , \; t \geq 0 \; , \; l = 1, 2, \ldots, L \, , \tag{4}$$

and assume *balancedness* at initialization, *i.e.*:

$$W_{l+1}(0)^\top W_{l+1}(0) = W_l(0) W_l(0)^\top \quad , \; l = 1, 2, \ldots, L-1 \, . \tag{5}$$

In particular, when considering random initialization, we assume that $\{W_l(0)\}_{l=1}^{L}$ are drawn from a joint probability distribution by which Equation (5) holds almost surely. This is an idealization of standard random near-zero initializations, *e.g.* Xavier ([31]) and He ([40]), by which Equation (5) holds approximately with high probability (note that the equation holds exactly in the standard "residual" setting of identity initialization — *cf.* [38, 10]). The condition of balanced initialization (Equation (5)) played an important role in the analysis of [6], facilitating derivation of a differential equation governing the product matrix of a linear neural network (see Lemma 4 in Subappendix G.2.1). It was shown in [6] empirically (and will be demonstrated again in Section 4) that there is an excellent match between the theoretical predictions of gradient flow with balanced initialization, and its practical realization via gradient descent with small learning rate and near-zero initialization. Other works (*e.g.* [7, 45]) have supported this match theoretically, and we provide additional support in Appendix D by extending our theory to the case of unbalanced initialization (Equation (5) holding approximately).

Formally stated, Conjecture 1 from [34] treats the case $L = 2$, where the product matrix $W_{L:1}$ (Equation (3)) holds $\alpha \cdot W_{init}$ at initialization, W_{init} being a fixed arbitrary full-rank matrix and α a varying positive scalar.[c] Taking time to infinity ($t \to \infty$) and then initialization size to zero ($\alpha \to 0^+$), the conjecture postulates that if the limit product matrix $\bar{W}_{L:1} := \lim_{\alpha\to 0^+}\lim_{t\to\infty} W_{L:1}$ exists and is a global optimum for the loss $\ell(\cdot)$ (Equation (1)), *i.e.* $\ell(\bar{W}_{L:1}) = 0$, then it will be a global optimum with minimal nuclear norm, meaning $\bar{W}_{L:1} \in \mathrm{argmin}_{W:\ell(W)=0} \|W\|_{nuclear}$. In contrast to Conjecture 1, Conjecture 2 from [8] can be interpreted as saying that for any depth $L \geq 2$ and any norm or quasi-norm $\|\cdot\|$, there exist observations $\{b_{i,j}\}_{(i,j)\in\Omega}$ for which global optimization of loss ($\lim_{\alpha\to 0^+}\lim_{t\to\infty} \ell(W_{1:L}) = 0$) does not imply minimization of $\|\cdot\|$ (*i.e.* we may have $\lim_{\alpha\to 0^+}\lim_{t\to\infty} \|W_{1:L}\| \neq \min_{W:\ell(W)=0}\|W\|$). Due to technical subtleties (for example the requirement of Conjecture 1 that a double limit of the product matrix with respect to time and initialization size exists), Conjectures 1 and 2 are not necessarily contradictory. However, they are in direct opposition in terms of the stances they represent — one supports the prospect of norms being able to explain implicit regularization in matrix factorization, and the other does not. The current paper seeks a resolution.

3 Implicit regularization can drive all norms to infinity

In this section we prove that for matrix factorization of depth $L \geq 2$, there exist observations $\{b_{i,j}\}_{(i,j)\in\Omega}$ with which optimizing the overparameterized objective (Equation (2)) via gradient flow (Equations (4) and (5)) leads — with probability 0.5 or more over random ("symmetric") initialization — *all* norms and quasi-norms of the product matrix (Equation (3)) to *grow towards infinity*, while its rank essentially decreases towards minimum. By this we not only affirm Conjecture 2, but in fact go beyond it in the following sense: *(i)* the conjecture allows chosen observations to depend on the norm or quasi-norm under consideration, while we show that the same set of observations can apply jointly to all norms and quasi-norms; and *(ii)* the conjecture requires norms and quasi-norms to be larger than minimal, while we establish growth towards infinity.

For simplicity of presentation, the current section delivers our construction and analysis in the setting $d = d' = 2$ (*i.e.* 2-by-2 matrix completion) — extension to different dimensions is straightforward (see Appendix E). We begin (Subsection 3.1) by introducing our chosen observations $\{b_{i,j}\}_{(i,j)\in\Omega}$ and discussing their properties. Subsequently (Subsection 3.2), we show that with these observations, decreasing loss often increases all norms and quasi-norms while lowering rank. Minimization of loss is treated thereafter (Subsection 3.3). Finally (Subsection 3.4), robustness of our construction to perturbations is established.

3.1 A simple matrix completion problem

Consider the problem of completing a 2-by-2 matrix based on the following observations:

$$\Omega = \{(1,2),(2,1),(2,2)\} \quad , \quad b_{1,2} = 1 \; , \; b_{2,1} = 1 \; , \; b_{2,2} = 0 \, . \tag{6}$$

The solution set for this problem (*i.e.* the set of matrices obtaining zero loss) is:

$$\mathcal{S} = \left\{ W \in \mathbb{R}^{2,2} : (W)_{1,2} = 1, (W)_{2,1} = 1, (W)_{2,2} = 0 \right\} \, . \tag{7}$$

Proposition 1 below states that minimizing a norm or quasi-norm along $W \in \mathcal{S}$ requires confining $(W)_{1,1}$ to a bounded interval, which for Schatten-p (quasi-)norms (in particular for nuclear, Frobenius and spectral norms)[5] is simply the singleton $\{0\}$.

Proposition 1. *For any norm or quasi-norm over matrices $\|\cdot\|$ and any $\epsilon > 0$, there exists a bounded interval $I_{\|\cdot\|,\epsilon} \subset \mathbb{R}$ such that if $W \in \mathcal{S}$ is an ϵ-minimizer of $\|\cdot\|$ (i.e. $\|W\| \leq \inf_{W'\in\mathcal{S}} \|W'\| + \epsilon$) then necessarily $(W)_{1,1} \in I_{\|\cdot\|,\epsilon}$. If $\|\cdot\|$ is a Schatten-p (quasi-)norm, then in addition $W \in \mathcal{S}$ minimizes $\|\cdot\|$ (i.e. $\|W\| = \inf_{W'\in\mathcal{S}} \|W'\|$) if and only if $(W)_{1,1} = 0$.*

Proof sketch (for complete proof see Subappendix G.3). The (weakened) triangle inequality allows us to lower bound $\|\cdot\|$ by $|(W)_{1,1}|$ (up to multiplicative and additive constants). Thus, the set of $(W)_{1,1}$ values corresponding to ϵ-minimizers must be bounded. If $\|\cdot\|$ is a Schatten-p (quasi-)norm, a straightforward analysis shows it is monotonically increasing with respect to $|(W)_{1,1}|$, implying it is minimized if and only if $(W)_{1,1} = 0$. □

In addition to norms and quasi-norms, we are also interested in the evolution of rank throughout optimization of a deep matrix factorization. More specifically, we are interested in the prospect of rank being implicitly minimized, as demonstrated empirically in [34, 8]. The discrete nature of rank renders its direct analysis unfavorable from a dynamical perspective (the rank of a matrix implies little about its proximity to low-rank), thus we consider the following surrogate measures: *(i) effective rank* (Definition 1 below; from [74]) — a continuous extension of rank used for numerical analyses; and *(ii) distance from infimal rank* (Definition 2 below) — (Frobenius) distance from the minimal rank that a given set of matrices may approach. According to Proposition 2 below, these measures independently imply that, although all solutions to our matrix completion problem — *i.e.* all $W \in \mathcal{S}$ (see Equation (7)) — have rank 2, it is possible to essentially minimize the rank to 1 by taking $|(W)_{1,1}| \to \infty$. Recalling Proposition 1, we conclude that in our setting, there is a direct contradiction between minimizing norms or quasi-norms and minimizing rank — the former requires confinement to some bounded interval, whereas the latter demands divergence towards infinity. This is the critical feature of our construction, allowing us to deem whether the implicit regularization in deep matrix factorization favors norms (or quasi-norms) over rank or vice versa.

Definition 1 (from [74])**.** The *effective rank of a matrix* $0 \neq W \in \mathbb{R}^{d,d'}$ with singular values $\{\sigma_r(W)\}_{r=1}^{\min\{d,d'\}}$ is defined to be $\mathrm{erank}(W) := \exp\{H(\rho_1(W), \rho_2(W), \ldots, \rho_{\min\{d,d'\}}(W))\}$, where $\{\rho_r(W) := \sigma_r(W) / \sum_{r'=1}^{\min\{d,d'\}} \sigma_{r'}(W)\}_{r=1}^{\min\{d,d'\}}$ is a distribution induced by the singular values,

[5] For $p \in (0, \infty]$, the *Schatten-p (quasi-)norm* of a matrix $W \in \mathbb{R}^{d,d'}$ with singular values $\{\sigma_r(W)\}_{r=1}^{\min\{d,d'\}}$ is defined as $\left(\sum_{r=1}^{\min\{d,d'\}} \sigma_r^p(W)\right)^{1/p}$ if $p < \infty$ and as $\max\{\sigma(W)\}_{r=1}^{\min\{d,d'\}}$ if $p = \infty$. It is a norm if $p \geq 1$ and a quasi-norm if $p < 1$. Notable special cases are nuclear (trace), Frobenius and spectral norms, corresponding to $p = 1$, 2 and ∞ respectively.

and $H(\rho_1(W), \rho_2(W), \ldots, \rho_{\min\{d,d'\}}(W)) := -\sum_{r=1}^{\min\{d,d'\}} \rho_r(W) \cdot \ln \rho_r(W)$ is its (Shannon) entropy (by convention $0 \cdot \ln 0 = 0$).

Definition 2. For a matrix space $\mathbb{R}^{d,d'}$, we denote by $D(\mathcal{S}, \mathcal{S}')$ the (Frobenius) distance between two sets $\mathcal{S}, \mathcal{S}' \subset \mathbb{R}^{d,d'}$ (*i.e.* $D(\mathcal{S}, \mathcal{S}') := \inf\{\|W - W'\|_{Fro} : W \in \mathcal{S}, W' \in \mathcal{S}'\}$), by $D(W, \mathcal{S}')$ the distance between a matrix $W \in \mathbb{R}^{d,d'}$ and the set $\mathcal{S}'$ (*i.e.* $D(W, \mathcal{S}') := \inf\{\|W - W'\|_{Fro} : W' \in \mathcal{S}'\}$), and by $\mathcal{M}_r$, for $r = 0, 1, \ldots, \min\{d, d'\}$, the set of matrices with rank r or less (*i.e.* $\mathcal{M}_r := \{W \in \mathbb{R}^{d,d'} : \mathrm{rank}(W) \leq r\}$). The *infimal rank of the set* $\mathcal{S}$ — denoted $\mathrm{irank}(\mathcal{S})$ — is defined to be the minimal r such that $D(\mathcal{S}, \mathcal{M}_r) = 0$. The *distance of a matrix* $W \in \mathbb{R}^{d,d'}$ *from the infimal rank of* $\mathcal{S}$ is defined to be $D(W, \mathcal{M}_{\mathrm{irank}(\mathcal{S})})$.

Proposition 2. *The effective rank (Definition 1) takes the values* $(1, 2]$ *along* $\mathcal{S}$ *(Equation* (7)*). For* $W \in \mathcal{S}$*, it is maximized when* $(W)_{1,1} = 0$*, and monotonically decreases to* 1 *as* $|(W)_{1,1}|$ *grows. Correspondingly, the infimal rank (Definition 2) of* $\mathcal{S}$ *is* 1*, and the distance of* $W \in \mathcal{S}$ *from this infimal rank is maximized when* $(W)_{1,1} = 0$*, monotonically decreasing to* 0 *as* $|(W)_{1,1}|$ *grows.*

Proof sketch (for complete proof see Appendix G.4). Analyzing the singular values of $W \in \mathcal{S}$ — $\sigma_1(W) \geq \sigma_2(W) \geq 0$ — reveals that: *(i)* $\sigma_1(W)$ attains a minimal value of 1 when $(W)_{1,1} = 0$, monotonically increasing to ∞ as $|(W)_{1,1}|$ grows; and *(ii)* $\sigma_2(W)$ attains a maximal value of 1 when $(W)_{1,1} = 0$, monotonically decreasing to 0 as $|(W)_{1,1}|$ grows. The results for effective rank, infimal rank and distance from infimal rank readily follow from this characterization. □

3.2 Decreasing loss increases norms

Consider the process of solving our matrix completion problem (Subsection 3.1) with gradient flow over a depth $L \geq 2$ matrix factorization (Section 2). Theorem 1 below states that if the product matrix (Equation (3)) has positive determinant at initialization, lowering the loss leads norms and quasi-norms to increase, while the rank essentially decreases.

Theorem 1. *Suppose we complete the observations in Equation* (6) *by employing a depth* $L \geq 2$ *matrix factorization, i.e. by minimizing the overparameterized objective (Equation* (2)*)*[d] *via gradient flow (Equations* (4) *and* (5)*). Denote by* $W_{L:1}(t)$ *the product matrix (Equation* (3)*) at time* $t \geq 0$ *of optimization, and by* $\ell(t) := \ell(W_{L:1}(t))$ *the corresponding loss (Equation* (1)*). Assume that* $\det(W_{L:1}(0)) > 0$*. Then, for any norm or quasi-norm over matrices* $\|\cdot\|$*:*

$$\|W_{L:1}(t)\| \geq a_{\|\cdot\|} \cdot \frac{1}{\sqrt{\ell(t)}} - b_{\|\cdot\|} \quad , \; t \geq 0 \,, \tag{8}$$

where $b_{\|\cdot\|} := \max\{\sqrt{2} a_{\|\cdot\|}, 8c_{\|\cdot\|}^2 \max_{i,j \in \{1,2\}} \|\mathbf{e}_i \mathbf{e}_j^\top\|\}$*,* $a_{\|\cdot\|} := \|\mathbf{e}_1 \mathbf{e}_1^\top\| / (\sqrt{2} c_{\|\cdot\|})$*, the vectors* $\mathbf{e}_1, \mathbf{e}_2 \in \mathbb{R}^2$ *form the standard basis, and* $c_{\|\cdot\|} \geq 1$ *is a constant with which* $\|\cdot\|$ *satisfies the weakened triangle inequality (see Footnote 2). On the other hand:*

$$\mathrm{erank}(W_{L:1}(t)) \leq \inf\nolimits_{W' \in \mathcal{S}} \mathrm{erank}(W') + \tfrac{2\sqrt{12}}{\ln(2)} \cdot \sqrt{\ell(t)} \qquad , \; t \geq 0 \,, \tag{9}$$

$$D(W_{L:1}(t), \mathcal{M}_{\mathrm{irank}(\mathcal{S})}) \leq 3\sqrt{2} \cdot \sqrt{\ell(t)} \qquad , \; t \geq 0 \,, \tag{10}$$

where $\mathrm{erank}(\cdot)$ *stands for effective rank (Definition 1), and* $D(\cdot\,, \mathcal{M}_{\mathrm{irank}(\mathcal{S})})$ *represents distance from the infimal rank (Definition 2) of the solution set* $\mathcal{S}$ *(Equation* (7)*).*

Proof sketch (for complete proof see Subappendix G.5). Using a dynamical characterization from [8] for the singular values of the product matrix (restated in Subappendix G.2.1 as Lemma 5), we show that the latter's determinant does not change sign, *i.e.* it remains positive. This allows us to lower bound $|(W_{L:1})_{1,1}(t)|$ by $1/\sqrt{\ell(t)}$ (up to multiplicative and additive constants). Relating $|(W_{L:1})_{1,1}(t)|$ to (quasi-)norms, effective rank and distance from infimal rank then leads to the desired bounds. □

An immediate consequence of Theorem 1 is that, if the product matrix (Equation (3)) has positive determinant at initialization, convergence to zero loss leads *all* norms and quasi-norms to *grow to infinity*, while the rank is essentially minimized. This is formalized in Corollary 1 below.

Corollary 1. *Under the conditions of Theorem 1, global optimization of loss, i.e.* $\lim_{t \to \infty} \ell(t) = 0$*, implies that for any norm or quasi-norm over matrices* $\|\cdot\|$ *we have that* $\lim_{t \to \infty} \|W_{L:1}(t)\| = \infty$*, where* $W_{L:1}(t)$ *is the product matrix of the deep factorization (Equation* (3)*) at time* t *of optimization. On the other hand:* $\lim_{t \to \infty} \mathrm{erank}(W_{L:1}(t)) = \inf_{W' \in \mathcal{S}} \mathrm{erank}(W')$ *and*

$\lim_{t\to\infty} D(W_{L:1}(t), \mathcal{M}_{\mathrm{irank}(\mathcal{S})}) = 0$*, where* $\mathrm{erank}(\cdot)$ *stands for effective rank (Definition 1), and* $D(\cdot\,, \mathcal{M}_{\mathrm{irank}(\mathcal{S})})$ *represents distance from the infimal rank (Definition 2) of the solution set* $\mathcal{S}$ *(Equation* (7)*).*

Proof. Taking the limit $\ell(t) \to 0$ in the bounds given by Theorem 1 establishes the results. □

Theorem 1 and Corollary 1 imply that in our setting (Subsection 3.1), where minimizing norms (or quasi-norms) and minimizing rank contradict each other, the implicit regularization of deep matrix factorization is willing to completely give up on the former in favor of the latter, at least on the condition that the product matrix (Equation (3)) has positive determinant at initialization. How probable is this condition? By Proposition 3 below, it holds with probability 0.5 if the product matrix is initialized by any one of a wide array of common distributions, including matrix Gaussian distribution with zero mean and independent entries, and a product of such. We note that rescaling (multiplying by $\alpha > 0$) initialization does not change the sign of product matrix's determinant, therefore as postulated by Conjecture 2, initialization close to the origin (along with small learning rate[e]) can *not* ensure convergence to solution with minimal norm or quasi-norm.

Proposition 3. *If* $W \in \mathbb{R}^{d,d}$ *is a random matrix whose entries are drawn independently from continuous distributions, each symmetric about the origin, then* $\Pr(\det(W) > 0) = \Pr(\det(W) < 0) = 0.5$*. Furthermore, for* $L \in \mathbb{N}$*, if* $W_1, W_2, \ldots, W_L \in \mathbb{R}^{d,d}$ *are random matrices drawn independently from continuous distributions, and there exists* $l \in \{1, 2, \ldots, L\}$ *with* $\Pr(\det(W_l) > 0) = 0.5$*, then* $\Pr(\det(W_L W_{L-1} \cdots W_1) > 0) = \Pr(\det(W_L W_{L-1} \cdots W_1) < 0) = 0.5$.

Proof sketch (for complete proof see Subappendix G.6). Multiplying a row of W by -1 keeps its distribution intact while flipping the sign of its determinant. This implies $\Pr(\det(W) > 0) = \Pr(\det(W) < 0)$. The first result then follows from the fact that a matrix drawn from a continuous distribution is almost surely non-singular. The second result is an outcome of the same fact, as well as the multiplicativity of determinant and the law of total probability. □

3.3 Convergence to zero loss

It is customary in the theory of deep learning (*cf.* [34, 36, 8]) to distinguish between implicit regularization — which concerns the type of solutions found in training — and the complementary question of whether training loss is globally optimized. We supplement our implicit regularization analysis (Subsection 3.2) by addressing this complementary question in two ways: *(i)* in Section 4 we empirically demonstrate that on the matrix completion problem we analyze (Subsection 3.1), gradient descent over deep matrix factorizations (Section 2) indeed drives training loss towards global optimum, *i.e.* towards zero; and *(ii)* in Proposition 4 below we theoretically establish convergence to zero loss for the special case of depth 2 and scaled identity initialization (treatment of additional depths and initialization schemes is left for future work). We note that when combined with Corollary 1, Proposition 4 affirms that in the latter special case, all norms and quasi-norms indeed grow to infinity while rank is essentially minimized.[f]

Proposition 4. *Consider the setting of Theorem 1 in the special case of depth* $L = 2$ *and initial product matrix (Equation* (3)*)* $W_{L:1}(0) = \alpha \cdot I$*, where* I *stands for the identity matrix and* $\alpha \in (0, 1]$*. Under these conditions* $\lim_{t\to\infty} \ell(t) = 0$*, i.e. the training loss is globally optimized.*

Proof sketch (for complete proof see Subappendix G.7). We first establish that the product matrix is positive definite for all t. This simplifies a dynamical characterization from [6] (restated as Lemma 4 in Subappendix G.2), yielding lucid differential equations governing the entries of the product matrix. Careful analysis of these equations then completes the proof. □

3.4 Robustness to perturbations

Our analysis (Subsection 3.2) has shown that when applying a deep matrix factorization (Section 2) to the matrix completion problem defined in Subsection 3.1, if the product matrix (Equation (3)) has positive determinant at initialization — a condition that holds with probability 0.5 under the wide variety of random distributions specified by Proposition 3 — then the implicit regularization drives *all* norms and quasi-norms *towards infinity*, while rank is essentially driven towards its minimum. A natural question is how common this phenomenon is, and in particular, to what extent does it persist if the observed entries we defined (Equation (6)) are perturbed. Theorem 2 in Appendix C generalizes Theorem 1 (from Subsection 3.2) to the case of arbitrary non-zero values for the off-diagonal observations $b_{1,2}, b_{2,1}$, and an arbitrary value for the diagonal observation $b_{2,2}$. In this generalization, the assumption (from Theorem 1) of the product matrix's determinant at initialization being positive

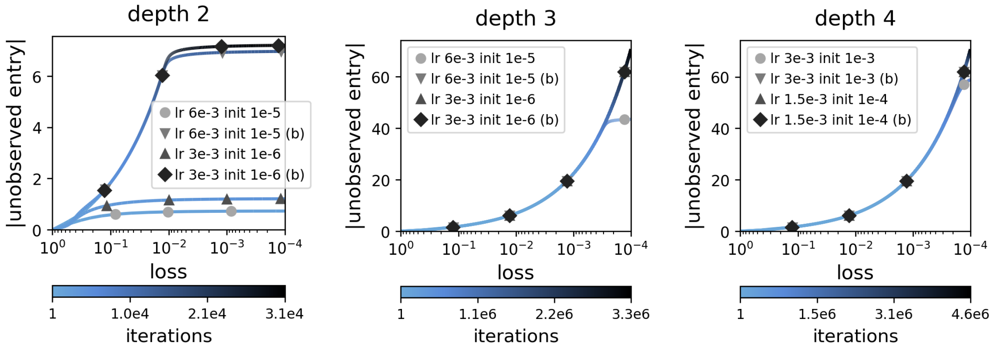

Figure 1: Implicit regularization in matrix factorization can drive *all* norms (and quasi-norms) *towards infinity*. For the matrix completion problem defined in Subsection 3.1, our analysis (Subsection 3.2) implies that with small learning rate and initialization close to the origin, when the product matrix (Equation (3)) is initialized to have positive determinant, gradient descent on a matrix factorization leads absolute value of unobserved entry to increase (which in turn means norms and quasi-norms increase) as loss decreases, *i.e.* as observations are fit. This is demonstrated in the plots above, which for representative runs, show absolute value of unobserved entry as a function of the loss (Equation 1), with iteration number encoded by color. Each plot corresponds to a different depth for the matrix factorization, and presents runs with varying configurations of learning rate and initialization (abbreviated as "lr" and "init", respectively). Both balanced (Equation 5) and unbalanced (layer-wise independent) random initializations were evaluated (former is marked by "(b)"). Independently for each depth, runs were iteratively carried out, with both learning rate and standard deviation for initialization decreased after each run, until the point where further reduction did not yield a noticeable change (presented runs are those from the last iterations of this process). Notice that depth, balancedness, and small learning rate and initialization, all contribute to the examined effect (absolute value of unobserved entry increasing as loss decreases), with the transition from depth 2 to 3 or more being most significant. Notice also that all runs initially follow the same curve, differing from one another in the point at which they divert (enter a phase where examined effect is lesser). While a complete investigation of these phenomena is left for future work, we provide a partial theoretical explanation in Appendix D. For further implementation details, and similar experiments with different matrix dimensions, as well as perturbed and repositioned observations, see Appendix F.

is modified to an assumption of it having the same sign as $b_{1,2} \cdot b_{2,1}$ (the probability of which is also 0.5 under the random distributions covered by Proposition 3). Conditioned on the modified assumption, the smaller $|b_{2,2}|$ is compared to $|b_{1,2} \cdot b_{2,1}|$, the higher the implicit regularization is guaranteed to drive norms and quasi-norms, and the lower it is guaranteed to essentially drive the rank. Two immediate implications of Theorem 2 are: *(i)* if the diagonal observation is unperturbed ($b_{2,2} = 0$), the off-diagonal ones ($b_{1,2}, b_{2,1}$) can take on *any* non-zero values, and the phenomenon of implicit regularization driving norms and quasi-norms towards infinity (while essentially driving rank towards its minimum) will persist; and *(ii)* this phenomenon gracefully recedes as the diagonal observation is perturbed away from zero. We note that Theorem 2 applies even if the unobserved entry is repositioned, thus our construction is robust not only to perturbations in observed values, but also to an arbitrary change in the observed locations. See Subappendix F.1 for empirical demonstrations.

4 Experiments

This section presents our empirical evaluations. We begin in Subsection 4.1 with deep matrix factorization (Section 2) applied to the settings we analyzed (Section 3). Then, we turn to Subsection 4.2 and experiment with an extension to tensor (multi-dimensional array) factorization. For brevity, many details behind our implementation, as well as some experiments, are deferred to Appendix F.

4.1 Analyzed settings

In [34], Gunasekar *et al.* experimented with matrix factorization, arriving at Conjecture 1. In the following work [8], Arora *et al.* empirically evaluated additional settings, ultimately arguing against Conjecture 1, and raising Conjecture 2. Our analysis (Section 3) affirmed Conjecture 2, by providing a setting in which gradient descent (with infinitesimally small learning rate and initialization arbitrarily close to the origin) over (shallow or deep) matrix factorization provably drives *all* norms (and quasi-norms) *towards infinity*. Specifically, we established that running gradient descent on the overparameterized matrix completion objective in Equation (2), where the observed entries are those defined in Equation (6), leads the unobserved entry to diverge to infinity as loss converges to zero. Figure 1 demonstrates this phenomenon empirically. Figures 4 and 5 in Subappendix F.1 extend the experiment by considering, respectively: different matrix dimensions (see Appendix E); and

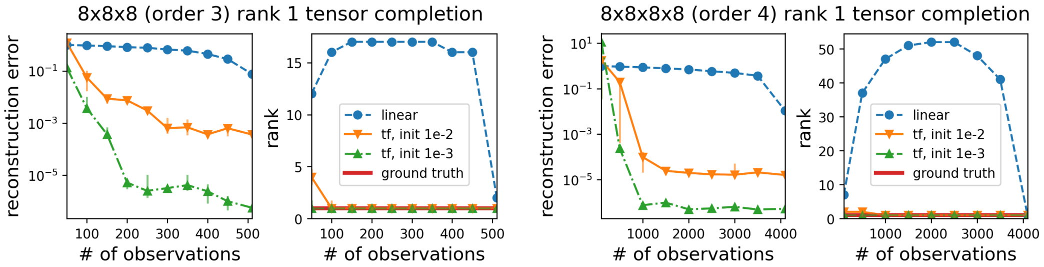

Figure 2: Gradient descent over tensor factorization exhibits an implicit regularization towards low (tensor) rank. Plots above report results of tensor completion experiments, comparing: *(i)* minimization of loss (Equation (11)) via gradient descent over tensor factorization (Equation (12) with R large enough for expressing any tensor) starting from (small) random initialization (method is abbreviated as "tf"); against *(ii)* trivial baseline that matches observations while holding zeros in unobserved locations — equivalent to minimizing loss via gradient descent over linear parameterization (*i.e.* directly over $\mathcal{W}$) starting from zero initialization (hence this method is referred to as "linear"). Each pair of plots corresponds to a randomly drawn low-rank ground truth tensor, from which multiple sets of observations varying in size were randomly chosen. The ground truth tensors corresponding to left and right pairs both have rank 1 (for results obtained with additional ground truth ranks see Figure 6 in Subappendix F.1), with sizes 8-by-8-by-8 (order 3) and 8-by-8-by-8-by-8 (order 4) respectively. The plots in each pair show reconstruction errors (Frobenius distance from ground truth) and ranks (numerically estimated) of final solutions as a function of the number of observations in the task, with error bars spanning interquartile range (25'th to 75'th percentiles) over multiple trials (differing in random seed for initialization), and markers showing median. For gradient descent over tensor factorization, we employed an adaptive learning rate scheme to reduce run times (see Subappendix F.2 for details), and iteratively ran with decreasing standard deviation for initialization, until the point at which further reduction did not yield a noticeable change (presented results are those from the last iterations of this process, with the corresponding standard deviations annotated by "init"). Notice that gradient descent over tensor factorization indeed exhibits an implicit tendency towards low rank (leading to accurate reconstruction of low-rank ground truth tensors), and that this tendency is stronger with smaller initialization. For further details and experiments see Appendix F.

perturbations and repositionings applied to observations (*cf.* Subsection 3.4). The figures confirm that the inability of norms (and quasi-norms) to explain implicit regularization in matrix factorization translates from theory to practice.

4.2 From matrix to tensor factorization

At the heart of our analysis (Section 3) lies a matrix completion problem whose solution set (Equation (7)) entails a direct contradiction between minimizing norms (or quasi-norms) and minimizing rank. We have shown that on this problem, gradient descent over (shallow or deep) matrix factorization is willing to completely give up on the former in favor of the latter. This suggests that, rather than viewing implicit regularization in matrix factorization through the lens of norms (or quasi-norms), a potentially more useful interpretation is *minimization of rank*. Indeed, while global minimization of rank is in the worst case computationally hard (*cf.* [73]), it has been shown in [8] (theoretically as well as empirically) that the dynamics of gradient descent over matrix factorization promote sparsity of singular values, and thus they may be interpreted as searching for low rank locally. As a step towards assessing the generality of this interpretation, we empirically explore an extension of matrix factorization to *tensor factorization*.[g]

In the context of matrix completion, (depth 2) matrix factorization amounts to optimizing the loss in Equation (1) by applying gradient descent to the parameterization $W = \sum_{r=1}^{R} \mathbf{w}_r \otimes \mathbf{w}'_r$, where $R \in \mathbb{N}$ is a predetermined constant, $\otimes$ stands for outer product,[6] and $\{\mathbf{w}_r \in \mathbb{R}^d\}_{r=1}^R$, $\{\mathbf{w}'_r \in \mathbb{R}^{d'}\}_{r=1}^R$ are the optimized parameters.[h] The minimal R required for this parameterization to be able to express a given $\bar{W} \in \mathbb{R}^{d,d'}$ is precisely the latter's rank. Implicit regularization towards low rank means that even when R is large enough for expressing any matrix (*i.e.* $R \geq \min\{d, d'\}$), solutions expressible (or approximable) with small R tend to be learned.

A generalization of the above is obtained by switching from matrices (tensors of *order* 2) to tensors of arbitrary order $N \in \mathbb{N}$. This gives rise to a *tensor completion* problem, with corresponding loss:

$$\ell : \mathbb{R}^{d_1, d_2, \ldots, d_N} \to \mathbb{R}_{\geq 0} \quad , \quad \ell(\mathcal{W}) = \frac{1}{2} \sum\nolimits_{(i_1, i_2, \ldots, i_N) \in \Omega} \left((\mathcal{W})_{i_1, i_2, \ldots, i_N} - b_{i_1, i_2, \ldots, i_N} \right)^2 , \quad (11)$$

[6] Given $\{\mathbf{v}^{(n)} \in \mathbb{R}^{d_n}\}_{n=1}^N$, the outer product $\mathbf{v}^{(1)} \otimes \mathbf{v}^{(2)} \otimes \cdots \otimes \mathbf{v}^{(N)} \in \mathbb{R}^{d_1, d_2, \ldots, d_N}$ — an order N tensor — is defined by $(\mathbf{v}^{(1)} \otimes \mathbf{v}^{(2)} \otimes \cdots \otimes \mathbf{v}^{(N)})_{i_1, i_2, \ldots, i_N} = (\mathbf{v}^{(1)})_{i_1} \cdot (\mathbf{v}^{(2)})_{i_2} \cdots (\mathbf{v}^{(N)})_{i_N}$.

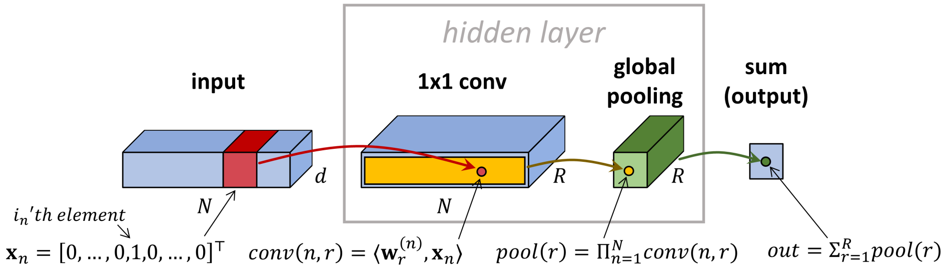
hidden layer
input
1x1 conv
global pooling
sum (output)
d
R
R
N
N
i_n'th element
$\mathbf{x}_n = [0, \dots, 0, 1, 0, \dots, 0]^\top$
$conv(n,r) = \langle \mathbf{w}_r^{(n)}, \mathbf{x}_n \rangle$
$pool(r) = \Pi_{n=1}^{N} conv(n,r)$
$out = \Sigma_{r=1}^{R} pool(r)$

Figure 3: Tensor factorizations correspond to convolutional arithmetic circuits (class of *non-linear* neural networks studied extensively), analogously to how matrix factorizations correspond to linear neural networks. Specifically, the tensor factorization in Equation (12) corresponds to the convolutional arithmetic circuit illustrated above (illustration assumes $d_1 = d_2 = \cdots = d_N = d$ to avoid clutter). The input to the network is a tuple $(i_1, i_2, \dots, i_N) \in \{1, 2, \dots, d_1\} \times \{1, 2, \dots, d_2\} \times \cdots \times \{1, 2, \dots, d_N\}$, represented via one-hot vectors $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \times \cdots \times \mathbb{R}^{d_N}$. These vectors are processed by a hidden layer comprising: *(i)* locally connected linear operator with R channels, the r'th one computing inner products against filters $(\mathbf{w}_r^{(1)}, \mathbf{w}_r^{(2)}, \dots, \mathbf{w}_r^{(N)}) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \times \cdots \times \mathbb{R}^{d_N}$ (this operator is referred to as "1×1 conv", appealing to the case of weight sharing, *i.e.* $\mathbf{w}_r^{(1)} = \mathbf{w}_r^{(2)} = \ldots = \mathbf{w}_r^{(N)}$); followed by *(ii)* global pooling computing products of all activations in each channel. The result of the hidden layer is then reduced through summation to a scalar — output of the network. Overall, given input tuple $(i_1, i_2, \dots, i_N)$, the network outputs $(\mathcal{W})_{i_1, i_2, \dots, i_N}$, where $\mathcal{W} \in \mathbb{R}^{d_1, d_2, \dots, d_N}$ is given by the tensor factorization in Equation (12). Notice that the number of terms (R) and the tunable parameters ($\{\mathbf{w}_r^{(n)}\}_{r,n}$) in the factorization respectively correspond to the width and the learnable filters of the network. Our tensor factorization (Equation (12)) was derived as an extension of a shallow (depth 2) matrix factorization, and accordingly, the convolutional arithmetic circuit it corresponds to is shallow (has a single hidden layer). Endowing the factorization with hierarchical structures would render it equivalent to *deep* convolutional arithmetic circuits (see [22] for details) — investigation of the implicit regularization in these models is viewed as a promising avenue for future research.

where $\{b_{i_1, i_2, \dots, i_N} \in \mathbb{R}\}_{(i_1, i_2, \dots, i_N) \in \Omega}$, $\Omega \subset \{1, 2, \dots, d_1\} \times \{1, 2, \dots, d_2\} \times \cdots \times \{1, 2, \dots, d_N\}$, stands for the set of observed entries. One may employ a tensor factorization by minimizing the loss in Equation (11) via gradient descent over the parameterization:

$$\mathcal{W} = \sum\nolimits_{r=1}^{R} \mathbf{w}_r^{(1)} \otimes \mathbf{w}_r^{(2)} \otimes \cdots \otimes \mathbf{w}_r^{(N)} \quad , \ \mathbf{w}_r^{(n)} \in \mathbb{R}^{d_n} \ , \ r = 1, 2, \dots, R \ , \ n = 1, 2, \dots, N \ , \quad (12)$$

where again, $R \in \mathbb{N}$ is a predetermined constant, $\otimes$ stands for outer product, and $\{\mathbf{w}_r^{(n)}\}_{r=1}^{R}{}_{n=1}^{N}$ are the optimized parameters.[i] In analogy with the matrix case, the minimal R required for this parameterization to be able to express a given $\bar{\mathcal{W}} \in \mathbb{R}^{d_1, d_2, \dots, d_N}$ is defined to be the latter's *(tensor) rank*. An implicit regularization towards low rank here would mean that even when R is large enough for expressing any tensor, solutions expressible (or approximable) with small R tend to be learned.

Figure 2 displays results of tensor completion experiments, in which tensor factorization (optimization of loss in Equation (11) via gradient descent over parameterization in Equation (12)) is applied to observations (*i.e.* $\{b_{i_1, i_2, \dots, i_N}\}_{(i_1, i_2, \dots, i_N) \in \Omega}$) drawn from a low-rank ground truth tensor. As can be seen in terms of both reconstruction error (distance from ground truth tensor) and (tensor) rank of the produced solutions, tensor factorizations indeed exhibit an implicit regularization towards low rank. The phenomenon thus goes beyond the special case of matrix (order 2 tensor) factorization. Theoretically supporting this finding is regarded as a promising direction for future research.

As discussed in Section 1, matrix completion can be seen as a prediction problem,[j] and matrix factorization as its solution with a *linear neural network*. In a similar vein, tensor completion may be viewed as a prediction problem, and tensor factorization as its solution with a *convolutional arithmetic circuit* — see Figure 3. Convolutional arithmetic circuits form a class of *non-linear* neural networks that has been studied extensively in theory (*cf.* [22, 19, 20, 23, 77, 54, 24, 9, 55]), and has also demonstrated promising results in practice (see [18, 21, 78]). Analogously to how the input-output mapping of a linear neural network is naturally represented by a matrix, that of a convolutional arithmetic circuit admits a natural representation as a tensor. Our experiments (Figure 2 and Figure 6 in Subappendix F.1) show that (at least in some settings) when learned via gradient descent, this tensor tends to have low rank. We thus obtain a second exemplar of a neural network architecture whose implicit regularization strives to lower a notion of rank for its input-output mapping. This leads us to believe that the phenomenon may be general, and formalizing notions of rank for input-output mappings of contemporary models may be key to explaining generalization in deep learning.

Broader Impact

The application of deep learning in practice is based primarily on trial and error, conventional wisdom and intuition, often leading to suboptimal performance, as well as compromise in important aspects such as safety, privacy and fairness. Developing rigorous theoretical foundations behind deep learning may facilitate a more principled use of the technology, alleviating aforementioned shortcomings. The current paper takes a step along this vein, by addressing the central question of implicit regularization induced by gradient-based optimization. While theoretical advances — particularly those concerned with explaining widely observed empirical phenomena — oftentimes do not pose apparent societal threats, a potential risk they introduce is misinterpretation by scientific readership. We have therefore made utmost efforts to present our results as transparently as possible.

Acknowledgments and Disclosure of Funding

This work was supported by Len Blavatnik and the Blavatnik Family foundation, as well as the Yandex Initiative in Machine Learning. The authors thank Nathan Srebro and Jason D. Lee for their illuminating comments which helped improve the manuscript.

References

[1] Evrim Acar, Daniel M Dunlavy, Tamara G Kolda, and Morten Mørup. Scalable tensor factorizations for incomplete data. *Chemometrics and Intelligent Laboratory Systems*, 106(1):41–56, 2011.

[2] Madhu S Advani and Andrew M Saxe. High-dimensional dynamics of generalization error in neural networks. *arXiv preprint arXiv:1710.03667*, 2017.

[3] Alnur Ali, Edgar Dobriban, and Ryan J Tibshirani. The implicit regularization of stochastic gradient flow for least squares. In *International Conference on Machine Learning (ICML)*, 2020.

[4] Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15:2773–2832, 2014.

[5] Raman Arora, Peter Bartlett, Poorya Mianjy, and Nathan Srebro. Dropout: Explicit forms and capacity control. *arXiv preprint arXiv:2003.03397*, 2020.

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

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

[8] Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 7413–7424, 2019.

[9] Emilio Rafael Balda, Arash Behboodi, and Rudolf Mathar. A tensor analysis on dense connectivity via convolutional arithmetic circuits. 2018.

[10] Peter Bartlett, Dave Helmbold, and Phil Long. Gradient descent with identity initialization efficiently learns positive definite linear transformations. In *International Conference on Machine Learning (ICML)*, pages 520–529, 2018.

[11] Mohamed Ali Belabbas. On implicit regularization: Morse functions and applications to matrix factorization. *arXiv preprint arXiv:2001.04264*, 2020.

[12] Alon Brutzkus and Amir Globerson. On the inductive bias of a cnn for orthogonal patterns distributions. *arXiv preprint arXiv:2002.09781*, 2020.

[13] Samuel Burer and Renato DC Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

[14] Changxiao Cai, Gen Li, H Vincent Poor, and Yuxin Chen. Nonconvex low-rank tensor completion from noisy data. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1863–1874, 2019.

[15] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717, 2009.

[16] Yuejie Chi, Yue M Lu, and Yuxin Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019.

[17] Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on Learning Theory (COLT)*, pages 1305–1338, 2020.

[18] Nadav Cohen and Amnon Shashua. Simnets: A generalization of convolutional networks. *Advances in Neural Information Processing Systems (NeurIPS), Deep Learning Workshop*, 2014.

[19] Nadav Cohen and Amnon Shashua. Convolutional rectifier networks as generalized tensor decompositions. *International Conference on Machine Learning (ICML)*, 2016.

[20] Nadav Cohen and Amnon Shashua. Inductive bias of deep convolutional networks through pooling geometry. *International Conference on Learning Representations (ICLR)*, 2017.

[21] Nadav Cohen, Or Sharir, and Amnon Shashua. Deep simnets. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[22] Nadav Cohen, Or Sharir, and Amnon Shashua. On the expressive power of deep learning: A tensor analysis. *Conference On Learning Theory (COLT)*, 2016.

[23] Nadav Cohen, Or Sharir, Yoav Levine, Ronen Tamari, David Yakira, and Amnon Shashua. Analysis and design of convolutional networks via hierarchical tensor decompositions. *Intel Collaborative Research Institute for Computational Intelligence (ICRI-CI) Special Issue on Deep Learning Theory*, 2017.

[24] Nadav Cohen, Ronen Tamari, and Amnon Shashua. Boosting dilated convolutional networks with mixed tensor decompositions. *International Conference on Learning Representations (ICLR)*, 2018.

[25] Assaf Dauber, Meir Feder, Tomer Koren, and Roi Livni. Can implicit bias explain generalization? stochastic convex optimization as a case study. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[26] Mark A Davenport and Justin Romberg. An overview of low-rank matrix recovery from incomplete observations. *IEEE Journal of Selected Topics in Signal Processing*, 10(4):608–622, 2016.

[27] Simon S Du, Wei Hu, and Jason D Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 384–395, 2018.

[28] Kelly Geyer, Anastasios Kyrillidis, and Amir Kalev. Low-rank regularization and solution uniqueness in over-parameterized matrix sensing. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pages 930–940, 2020.

[29] Gauthier Gidel, Francis Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in linear neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3196–3206, 2019.

[30] Daniel Gissin, Shai Shalev-Shwartz, and Amit Daniely. The implicit bias of depth: How incremental learning drives generalization. *International Conference on Learning Representations (ICLR)*, 2020.

[31] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.

[32] Sebastian Goldt, Madhu Advani, Andrew M Saxe, Florent Krzakala, and Lenka Zdeborová. Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6979–6989, 2019.

[33] Gene H Golub and Charles F Van Loan. *Matrix computations*, volume 3. JHU press, 2012.

[34] Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6151–6159, 2017.

[35] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80, pages 1832–1841, 2018.

[36] Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 9461–9471, 2018.

[37] Wolfgang Hackbusch. *Tensor spaces and numerical tensor calculus*, volume 42. Springer, 2012.

[38] Moritz Hardt and Tengyu Ma. Identity matters in deep learning. *International Conference on Learning Representations (ICLR)*, 2016.

[39] Johan Håstad. Tensor rank is np-complete. *Journal of algorithms (Print)*, 11(4):644–654, 1990.

[40] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[41] Yulij Ilyashenko and Sergei Yakovenko. *Lectures on analytic differential equations*, volume 86. American Mathematical Soc., 2008.

[42] Ilse CF Ipsen and Rizwana Rehman. Perturbation bounds for determinants and characteristic polynomials. *SIAM Journal on Matrix Analysis and Applications*, 30(2):762–776, 2008.

[43] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems (NeurIPS)*, pages 8571–8580, 2018.

[44] Prateek Jain and Sewoong Oh. Provable tensor factorization with missing data. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1431–1439, 2014.

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

[46] Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Conference on Learning Theory (COLT)*, pages 1772–1798, 2019.

[47] Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[48] Dimitris Kalimeris, Gal Kaplun, Preetum Nakkiran, Benjamin Edelman, Tristan Yang, Boaz Barak, and Haofeng Zhang. Sgd on neural networks learns functions of increasing complexity. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3491–3501, 2019.

[49] Lars Karlsson, Daniel Kressner, and André Uschmajew. Parallel algorithms for tensor completion in the cp format. *Parallel Computing*, 57:222–234, 2016.

[50] Tosio Kato. *Perturbation theory for linear operators*, volume 132. Springer Science & Business Media, 2013.

[51] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

[52] Steven G Krantz and Harold R Parks. *A primer of real analytic functions*. Springer Science & Business Media, 2002.

[53] Andrew K Lampinen and Surya Ganguli. An analytic theory of generalization dynamics and transfer learning in deep linear networks. *International Conference on Learning Representations (ICLR)*, 2019.

[54] Yoav Levine, David Yakira, Nadav Cohen, and Amnon Shashua. Deep learning and quantum entanglement: Fundamental connections with implications to network design. *International Conference on Learning Representations (ICLR)*, 2018.

[55] Yoav Levine, Or Sharir, Nadav Cohen, and Amnon Shashua. Quantum entanglement in deep learning architectures. *To appear in Physical Review Letters*, 2019.

[56] Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Proceedings of the 31st Conference On Learning Theory (COLT)*, pages 2–47, 2018.

[57] Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. *International Conference on Learning Representations (ICLR)*, 2020.

[58] Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion. In *International Conference on Machine Learning (ICML)*, pages 3351–3360, 2018.

[59] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on Learning Theory (COLT)*, pages 2388–2464, 2019.

[60] Rotem Mulayoff and Tomer Michaeli. Unique properties of wide minima in deep networks. In *International Conference on Machine Learning (ICML)*, 2020.

[61] Mor Shpigel Nacson, Suriya Gunasekar, Jason Lee, Nathan Srebro, and Daniel Soudry. Lexicographic and depth-sensitive margins in homogeneous and non-homogeneous deep models. In *International Conference on Machine Learning (ICML)*, pages 4683–4692, 2019.

[62] Mor Shpigel Nacson, Jason Lee, Suriya Gunasekar, Pedro Henrique Pamplona Savarese, Nathan Srebro, and Daniel Soudry. Convergence of gradient descent on separable data. In *Proceedings of Machine Learning Research*, volume 89, pages 3420–3428, 2019.

[63] Atsuhiro Narita, Kohei Hayashi, Ryota Tomioka, and Hisashi Kashima. Tensor factorization using auxiliary information. *Data Mining and Knowledge Discovery*, 25(2):298–324, 2012.

[64] Behnam Neyshabur. Implicit regularization in deep learning. *PhD thesis*, 2017.

[65] 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.

[66] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5947–5956, 2017.

[67] Samet Oymak and Mahdi Soltanolkotabi. Overparameterized nonlinear learning: Gradient descent takes the shortest path? In *International Conference on Machine Learning (ICML)*, pages 4951–4960, 2019.

[68] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.

[69] Robert T Powers and Erling Størmer. Free states of the canonical anticommutation relations. *Communications in Mathematical Physics*, 16(1):1–33, 1970.

[70] Adityanarayanan Radhakrishnan, Eshaan Nichani, Daniel Bernstein, and Caroline Uhler. Balancedness and alignment are unlikely in linear neural networks. *arXiv preprint arXiv:2003.06340*, 2020.

[71] Nasim Rahaman, Devansh Arpit, Aristide Baratin, Felix Draxler, Min Lin, Fred A Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of deep neural networks. In *International Conference on Machine Learning (ICML)*, pages 5301–5310, 2019.

[72] Noam Razin and Nadav Cohen. Implicit regularization in deep learning may not be explainable by norms. *arXiv preprint arXiv:2005.06398*, 2020.

[73] Benjamin Recht, Weiyu Xu, and Babak Hassibi. Null space conditions and thresholds for rank minimization. *Mathematical programming*, 127(1):175–202, 2011.

[74] Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European Signal Processing Conference*, pages 606–610. IEEE, 2007.

[75] Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *International Conference on Learning Representations (ICLR)*, 2014.

[76] Vatsal Shah, Anastasios Kyrillidis, and Sujay Sanghavi. Minimum weight norm models do not always generalize well for over-parameterized problems. *arXiv preprint arXiv:1811.07055*, 2018.

[77] Or Sharir and Amnon Shashua. On the expressive power of overlapping architectures of deep learning. *International Conference on Learning Representations (ICLR)*, 2018.

[78] Or Sharir, Ronen Tamari, Nadav Cohen, and Amnon Shashua. Tensorial mixture models. *arXiv preprint*, 2016.

[79] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

[80] Arun Suggala, Adarsh Prasad, and Pradeep K Ravikumar. Connecting optimization and regularization paths. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 10608–10619, 2018.

[81] Gerald Teschl. *Ordinary differential equations and dynamical systems*, volume 140. American Mathematical Soc., 2012.

[82] Stephen Tu, Ross Boczar, Max Simchowitz, Mahdi Soltanolkotabi, and Ben Recht. Low-rank solutions of linear matrix equations via procrustes flow. In *International Conference on Machine Learning (ICML)*, pages 964–973, 2016.

[83] Colin Wei, Sham Kakade, and Tengyu Ma. The implicit and explicit regularization effects of dropout. In *International Conference on Machine Learning (ICML)*, 2020.

[84] Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory (COLT)*, pages 3635–3673, 2020.

[85] Xiaoxia Wu, Edgar Dobriban, Tongzheng Ren, Shanshan Wu, Zhiyuan Li, Suriya Gunasekar, Rachel Ward, and Qiang Liu. Implicit regularization of normalization methods. *arXiv preprint arXiv:1911.07956*, 2019.

[86] Dong Xia and Ming Yuan. On polynomial time methods for exact low rank tensor completion. *arXiv preprint arXiv:1702.06980*, 2017.

[87] Tatsuya Yokota, Qibin Zhao, and Andrzej Cichocki. Smooth parafac decomposition for tensor completion. *IEEE Transactions on Signal Processing*, 64(20):5423–5436, 2016.

[88] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *International Conference on Learning Representations (ICLR)*, 2017.

[89] Pan Zhou, Canyi Lu, Zhouchen Lin, and Chao Zhang. Tensor factorization for low-rank tensor completion. *IEEE Transactions on Image Processing*, 27(3):1152–1163, 2017.

Notes

[a] The *nuclear norm* (also known as *trace norm*) of a matrix is the sum of its singular values, regarded as a convex relaxation of rank.

[b] A technical subtlety of optimization in continuous time is that in principle, it is possible to asymptote (diverge to infinity) after finite time. In such a case, the asymptote is regarded as the end of optimization, and time tending to infinity ($t \to \infty$) is to be interpreted as tending towards that point.

[c] The formal statement in [34] applies to symmetric matrix factorization and positive definite W_{init}, but it is claimed thereafter that affirming the conjecture would imply the same for the asymmetric setting considered in this paper. We also note that the conjecture is stated in the context of matrix sensing, thus in particular applies to matrix completion (a special case).

[d] As stated in Section 2, we consider full-dimensional factorizations, in this case meaning that hidden dimensions $d_1, d_2, \ldots, d_{L-1}$ are all greater than or equal to 2.

[e] Recall that gradient flow corresponds to gradient descent with infinitesimally small learning rate.

[f] Notice that under (positively) scaled identity initialization the determinant of the product matrix (Equation (3)) is positive, as required by Corollary 1.

[g] The reader is referred to [51] and [37] for an introduction to tensor factorizations.

[h] To see that this parameterization is equivalent to the usual form $W = W_2 W_1$, simply view R as the dimension shared between W_1 and W_2, $\{\mathbf{w}_r\}_{r=1}^R$ as the columns of W_2, and $\{\mathbf{w}'_r\}_{r=1}^R$ as the rows of W_1.

[i] There exist many types of tensor factorizations (*cf.* [51, 37]). We treat here the classic and most basic one, known as *CANDECOMP/PARAFAC (CP)*.

[j] Observed entries in the matrix to recover stand for training examples, and unobserved entries for test set.

A Summary

The extent to which norms (and quasi-norms) can explain the implicit regularization induced by gradient-based optimization is a central question in the theory of deep learning. A standard test-bed for its study is matrix factorization — matrix completion via linear neural networks trained by gradient descent — which in practice tends to produce low-rank solutions. It is an open problem whether the implicit regularization in matrix factorization can be characterized as minimization of a norm (or quasi-norm) — Conjecture 1 from [34] supports this supposition, whereas Conjecture 2 from [8] opposes it. We presented a simple (and robust to perturbations) matrix completion setting for which, with probability 0.5 or more over random initialization of gradient descent, the implicit regularization in matrix factorization provably drives *all* norms (and quasi-norms) to *grow towards infinity*, while rank is essentially minimized. This affirms Conjecture 2, and although it does not formally refute Conjecture 1 (the latter's technical assumptions are not necessarily satisfied by our setting), we believe that in essence our result implies that norm (or quasi-norm) minimization cannot explain implicit regularization in matrix factorization, let alone in deep learning altogether.

The crux behind the matrix completion setting we defined is that its solution set entails a direct contradiction between minimizing norms (or quasi-norms) and minimizing rank. The fact that the former is given up on in favor of the latter suggests that, rather than viewing implicit regularization in matrix factorization through the lens of norms (or quasi-norms), a potentially more useful interpretation is minimization of rank. As a step towards assessing the generality of this interpretation, we experimented with an extension of matrix factorization to tensor factorization, and found that it too exhibits an implicit regularization towards low rank, where rank is defined in the context of tensors. Similarly to how matrix factorization corresponds to a linear neural network whose input-output mapping is represented by a matrix, tensor factorization corresponds to a convolutional arithmetic circuit (certain type of *non-linear* neural network) whose input-output mapping is represented by a tensor. We thus obtain a second exemplar of a neural network architecture whose implicit regularization strives to lower a notion of rank for its input-output mapping. Theoretical investigation of the implicit regularization in tensor factorization is regarded as a promising direction for future research. More broadly, we believe that neural networks minimizing notions of rank for their input-output mappings may be a general phenomenon, and hypothesize that formalizing such notions in the context of contemporary models may be key to explaining generalization in deep learning.

B Related work

Theoretical analysis of implicit regularization in deep learning is a highly active area of research. Our work extends the bulk of literature concerning mathematical characterization of the implicit regularization induced by gradient-based optimization.[7] Existing characterizations focus on different aspects of learning, for example: dynamics of optimization ([2, 29, 53, 8, 32, 48, 30]); curvature ("flatness") of obtained minima ([60]); frequency spectrum of learned input-output mappings ([71]); invariant quantities throughout training ([27]); and statistical properties imported from data ([12]). A ubiquitous approach, arguably more prevalent than the aforementioned, is to demonstrate that learned input-output mappings minimize some notion of norm, or analogously, maximize some notion of margin. Works along this line have treated various models,[8] including: linear (single-layer) predictors ([79, 35, 46, 3]); normalized linear models ([85]); certain polynomially parameterized linear models ([84]); homogeneous (and sum of homogeneous) models ([61, 57, 47]); ultra-wide neural networks ([43, 59, 17, 67]); linear neural networks with a single output ([62, 36, 45]); and matrix factorization — the subject of our inquiry.

Matrix factorization is perhaps the most extensively studied model in the context of implicit regularization induced by non-convex gradient-based optimization. It corresponds to linear neural networks with multiple inputs and outputs, typically trained to recover low-rank linear mappings. The literature on matrix factorization for low-rank matrix recovery is far too broad to cover here — we refer to [16] for a recent survey, while mentioning that the technique is often attributed to [13]. Notable works proving successful recovery of a low-rank matrix via matrix factorization trained by gradient descent with no explicit regularization are [82, 58, 56]. Of these, [56] can be viewed as affirming Conjecture 1 (from [34]) for a certain special case.[9] [11] has affirmed Conjecture 1 under different assumptions, but nonetheless argued empirically that it does not hold true in general, resonating with Conjecture 2 (from [8]). To the best of our knowledge, no theoretical support for the latter was provided prior to its proof in this paper. We note that the proof relies on technical results derived in [6] and [8] (restated in Subappendix G.2.1 for completeness).

Extending the research on matrix factorization, the use of tensor factorization for recovering low-rank tensors is a frequent topic of investigation (*cf.* [14, 86, 89, 87, 49, 63, 44, 1, 4]). Nevertheless, the experiments reported in this paper provide the first evidence we are aware of for such use to be successful under gradient-based optimization with no explicit regularization (in particular without imposing low-rank on the tensor factorization).

C Robustness to perturbations theorem

Theorem 2. *Consider the setting of Theorem 1 subject to the following changes:* (i) *the observations from Equation* (6) *are generalized to:*

$$\Omega = \{(1,2),(2,1),(2,2)\} \quad , \quad b_{1,2} = z \in \mathbb{R}\backslash\{0\} \,,\; b_{2,1} = z' \in \mathbb{R}\backslash\{0\} \,,\; b_{2,2} = \epsilon \in \mathbb{R} \,, \tag{13}$$

leading to the following solution set in place of that from Equation (7)*:*

$$\widetilde{\mathcal{S}} = \left\{ W \in \mathbb{R}^{2,2} : (W)_{1,2} = z, (W)_{2,1} = z', (W)_{2,2} = \epsilon \right\} \;; \tag{14}$$

and (ii) *the assumption* $\det(W_{L:1}(0)) > 0$ *is generalized to* $\mathrm{sign}(\det(W_{L:1}(0))) = \mathrm{sign}(z \cdot z')$*, where* $W_{L:1}(t)$ *denotes the product matrix (Equation* (3)*) at time* $t \geq 0$ *of optimization. Under these conditions, for any norm or quasi-norm over matrices* $\|\cdot\|$*:*

$$\|W_{L:1}(t)\| \geq a_{\|\cdot\|} \cdot \frac{|z| \cdot |z'|}{|\epsilon| + \sqrt{2\ell(t)}} - b_{\|\cdot\|} \quad , \; t \geq 0 \,, \tag{15}$$

where $b_{\|\cdot\|} := \max\left\{ {}^{a_{\|\cdot\|} \cdot |z| \cdot |z'|}/{}_{(|\epsilon| + \min\{|z|,|z'|\})} \,,\, 8c^2_{\|\cdot\|} \max\{|z|\,,|z'|, |\epsilon|\} \max_{i,j \in \{1,2\}} \left\| \mathbf{e}_i \mathbf{e}_j^\top \right\| \right\}$*,* $a_{\|\cdot\|} := \left\| \mathbf{e}_1 \mathbf{e}_1^\top \right\| / c_{\|\cdot\|}$*, the vectors* $\mathbf{e}_1, \mathbf{e}_2 \in \mathbb{R}^2$ *form the standard basis, and* $c_{\|\cdot\|} \geq 1$ *is a constant*

[7] As opposed to works studying the relation between implicit regularization and generalization (*cf.* [64, 66, 76]), or ones analyzing other sources of implicit regularization such as dropout (*e.g.* [83, 5]).

[8] Limiting such treatments to particular models is necessary, as in general one can not expect a gradient-based optimizer to yield minimal norm solutions over all possible objectives. Indeed, [80] and [25] have shown that there exist (carefully crafted) objectives over which variants of gradient descent do not produce minimal norm solutions. This accords with the conventional wisdom by which implicit regularization does not stem from an optimizer alone, but from its combination with a model (class of objectives).

[9] For a case related to (yet different from) that of [56], it was shown in [28] that the set of matrices fitting observations is in fact a singleton, meaning Conjecture 1 holds trivially.

with which $\|\cdot\|$ satisfies the weakened triangle inequality (see Footnote 2). On the other hand:

$$\mathrm{erank}(W_{L:1}(t)) \leq \inf\nolimits_{W' \in \widetilde{\mathcal{S}}} \mathrm{erank}(W') + \tfrac{16}{\min\{|z|,|z'|\}} \Big(|\epsilon| + \sqrt{2\ell(t)} \Big) \qquad , \; t \geq 0 \,, \tag{16}$$

$$D(W_{L:1}(t), \mathcal{M}_{\mathrm{irank}(\widetilde{\mathcal{S}})}) \leq 4\,|\epsilon| + \Big(4 + \tfrac{\sqrt{|z| \cdot |z'|}}{\min\{|z|,|z'|\}} \Big) \sqrt{2\ell(t)} \qquad , \; t \geq 0 \,, \tag{17}$$

where $\mathrm{erank}(\cdot)$ *stands for effective rank (Definition 1), and* $D(\cdot\,, \mathcal{M}_{\mathrm{irank}(\widetilde{\mathcal{S}})})$ *represents distance from the infimal rank (Definition 2) of the solution set* $\widetilde{\mathcal{S}}$*. Moreover, Equations* (15)*,* (16) *and* (17) *hold even if the above setting is further generalized as follows:* (i) *the unobserved entry resides in location* $(i, j) \in \{1, 2\} \times \{1, 2\}$*, with* $z, z' \in \mathbb{R} \backslash \{0\}$ *observed in the adjacent locations and* $\epsilon \in \mathbb{R}$ *in the diagonally-opposite one; and* (ii) *the sign of* $\det(W_{L:1}(0))$ *is equal to that of* $z \cdot z'$ *if* $i = j$*, and opposite to it otherwise.*

Proof sketch (for complete proof see Subappendix G.8). The proof follows a line similar to that of Theorem 1, with slightly more involved derivations. □

Disqualifying implicit minimization of norms in finite settings Theorem 1 (in Subsection 3.2), which applies to our original (unperturbed) matrix completion problem (Subsection 3.1), has shown that the implicit regularization of deep matrix factorization (Section 2) does not minimize norms or quasi-norms, by establishing lower bounds (Equation (8) with $\|\cdot\|$ ranging over all possible norms and quasi-norms) that tend to infinity as the training loss $\ell(t)$ converges to zero. The more general Theorem 2 allows disqualifying implicit minimization of norms or quasi-norms without requiring divergence to infinity. To see this, consider the lower bounds it establishes (Equation (15) with $\|\cdot\|$ ranging over all norms and quasi-norms), and in particular their limits as $\ell(t) \rightarrow 0$. If the observed value ϵ is different from zero, these limits are all finite. Moreover, given a particular norm or quasi-norm $\|\cdot\|$, we may choose ϵ different from zero, yet small enough such that the lower bound for $\|\cdot\|$ has limit arbitrarily larger than the infimum of $\|\cdot\|$ over the solution set.

D Extension to unbalanced initialization

In this appendix we present extensions of our theory to cases where initialization is unbalanced, *i.e.* in which Equation (5) only holds approximately. For simplicity, we limit the presentation to the square setting, where all dimensions of the deep matrix factorization ($d_0, d_1, \ldots, d_L$; see Section 2) are equal to $d \in \mathbb{N}$.

The following definition quantifies unbalancedness.

Definition 3. The *unbalancedness magnitude* of matrices $\{W_l \in \mathbb{R}^{d,d}\}_{l=1}^{L}$ is defined to be:

$$\max\nolimits_{l \in \{1,2,\ldots,L-1\}} \left\| W_{l+1}^{\top} W_{l+1} - W_l W_l^{\top} \right\|_{nuclear} . \tag{18}$$

We will present two approaches for showing that approximate versions of our main theoretical results (Theorems 1 and 2) hold if unbalancedness magnitude at initialization is small: *(i)* using continuity of optimizer trajectory with respect to its initialization (Subappendix D.1); and *(ii)* employing conservation of unbalancedness magnitude throughout optimization (Subappendix D.2). Both approaches rely on the fact that small unbalancedness magnitude implies proximity to perfect balancedness, as stated formally in the following lemma.

Lemma 1. *For any matrices* $\{W_l \in \mathbb{R}^{d,d}\}_{l=1}^{L}$ *with unbalancedness magnitude (Equation* (18)*) equal to* ϵ*, there exist* $\{W'_l \in \mathbb{R}^{d,d}\}_{l=1}^{L}$ *that are balanced (i.e. have unbalancedness magnitude zero), such that* $\|W_l - W'_l\|_{Fro} \leq (l-1) \cdot \sqrt{\epsilon}$ *for all* $l = 1, 2, \ldots, L$*.*

Proof sketch (for complete proof see Subappendix G.9). Based on singular value decompositions of $W_1, W_2, \ldots, W_L$, the proof provides an explicit construction for $W'_1, W'_2, \ldots, W'_L$. Starting with $W'_1 := W_1$, for $l = 2, 3, \ldots, L$ the matrices W'_l are defined such that: *(i)* $W_l'^{\top} W'_l = W'_{l-1} W_{l-1}'^{\top}$; and *(ii)* $\|W_l - W'_l\|_{Fro} \leq \left\|W_{l-1} - W'_{l-1}\right\|_{Fro} + \sqrt{\epsilon}$. The former ensures that $W'_1, W'_2, \ldots, W'_L$ are balanced, whereas the latter implies $\|W_l - W'_l\|_{Fro} \leq (l-1) \cdot \sqrt{\epsilon}$ for all $l = 1, 2, \ldots, L$. □

D.1 First approach: continuity with respect to initialization

Trajectories of gradient flow over a smooth objective are Lipschitz continuous with respect to their initialization, in the sense that for any $T > 0$, the location at time T of optimization is a Lipschitz continuous function of the initial point. This fact is established by Lemma 2 below.

Lemma 2. *Let $\mathcal{D}$ be an open domain in Euclidean space, and let $f : \mathcal{D} \to \mathbb{R}$ be a twice continuously differentiable function. Denote by $\|\cdot\|_2$ the Euclidean norm, and assume that $f(\cdot)$ is β-smooth with respect to $\|\cdot\|_2$ for some $\beta \geq 0$.*[10] *Let $\theta, \theta' : [0, T] \to \mathcal{D}$, where $T > 0$, be two curves born from gradient flow over $f(\cdot)$:*

$$\theta(0) = \theta_0 \in \mathcal{D} \quad , \quad \dot{\theta}(t) := \tfrac{d}{dt}\theta(t) \ = -\nabla f(\theta(t)) \ , \ t \in [0, T] \, ,$$

$$\theta'(0) = \theta'_0 \in \mathcal{D} \quad , \quad \dot{\theta}'(t) := \tfrac{d}{dt}\theta'(t) = -\nabla f(\theta'(t)) \ , \ t \in [0, T] \, .$$

Then, for any $\bar{t} \in [0, T]$ it holds that:

$$\|\theta(\bar{t}) - \theta'(\bar{t})\|_2 \leq \|\theta(0) - \theta'(0)\|_2 \cdot \exp(\beta \cdot \bar{t}) \, . \tag{19}$$

Proof sketch (for complete proof see Subappendix G.10). Define the function $g : [0, T] \to \mathbb{R}_{\geq 0}$ by $g(t) := \|\theta(t) - \theta'(t)\|_2^2$. Since $f(\cdot)$ is β-smooth, it holds that $\dot{g}(t) := \frac{d}{dt}g(t) \leq 2\beta \cdot g(t)$ for all $t \in [0, T]$. Dividing the latter inequality by $g(t)$ (with special treatment for the case where $g(t) = 0$) and integrating over time yields the desired result. □

Specializing Lemma 2 to deep matrix factorization (Section 2) yields the following result.

Proposition 5. *Consider the overparameterized objective corresponding to a depth L matrix factorization applied to an arbitrary matrix completion task (see Equations (2) and (3)). Let $\theta(t) := (W_1(t), W_2(t), \ldots, W_L(t))$ and $\theta'(t) := (W'_1(t), W'_2(t), \ldots, W'_L(t))$ be two (arbitrary) curves born from gradient flow over this objective (cf. Equation (4)). Given $T > 0$, denote $R := \sup_{t \in [0,T]} \max\{\|\theta(t)\|_{Fro}, \|\theta'(t)\|_{Fro}\}$.*[11] *Then, for any $\bar{t} \in [0, T]$ it holds that:*

$$\|\theta(\bar{t}) - \theta'(\bar{t})\|_{Fro} \leq \|\theta(0) - \theta'(0)\|_{Fro} \cdot \exp\left(LR^{L-2}\left(2R^L + B\right) \cdot \bar{t}\right) \, , \tag{20}$$

where $B := (\sum_{(i,j) \in \Omega} b_{i,j}^2)^{0.5}$, with $\{b_{i,j}\}_{(i,j) \in \Omega}$ standing for the observed matrix entries.

Proof sketch (for complete proof see Subappendix G.11). The proof follows from Lemma 2, and the fact that for any $R' > 0$ the overparameterized objective is $LR'^{L-2}\left(2R'^L + B\right)$-smooth over $\mathcal{D}_{R'} := \{(W_1, W_2, \ldots, W_L) : \|(W_1, W_2, \ldots, W_L)\|_{Fro} < R'\}$. □

Combining Proposition 5 with Lemma 1 makes it possible to derive extensions of Theorems 1 and 2 in which the assumption of initialization being perfectly balanced (Equation (5)) is relaxed to a requirement for small unbalancedness magnitude (Definition 3). The underlying idea is as follows. An initialization with small unbalancedness magnitude is close to one which is balanced (Lemma 1), and for the latter Theorems 1 and 2 may be applied. The distance between gradient flow trajectories emanating from the two initializations is controlled (Proposition 5), therefore results of Theorems 1 and 2 (bounds on norms, quasi-norms, effective rank and distance from infimal rank) carry over — with additional error terms — to the trajectory originating from the unbalanced initialization. A drawback of this approach is that the bounds on distance between trajectories, and accordingly the error terms incurred, grow exponentially with time (see Equation (20)). In the next subappendix we present a different approach that takes into account specific properties of gradient flow over deep matrix factorization, allowing one to overcome this exponential growth (for depth $L \geq 3$).

D.2 Second approach: conservation of unbalancedness magnitude

Lemma 3 below shows that unbalancedness magnitude (Definition 3) is a conserved quantity of gradient flow over deep matrix factorization (Section 2).

[10] That is, for any $\theta_1, \theta_2 \in \mathcal{D}$ it holds that $\|\nabla f(\theta_2) - \nabla f(\theta_1)\|_2 \leq \beta \cdot \|\theta_2 - \theta_1\|_2$.

[11] The Frobenius norm of a matrix tuple is defined as the Euclidean norm of their concatenation as a vector, so for example $\|\theta(t)\|_{Fro} = (\sum_{l=1}^{L} \|W_l(t)\|_{Fro}^2)^{0.5}$.

Lemma 3. *Consider the overparameterized objective corresponding to a depth L matrix factorization applied to an arbitrary matrix completion task (see Equations* (2) *and* (3)*). Let $(W_1(t), W_2(t), \dots, W_L(t))$ be a curve born from gradient flow over this objective (cf. Equation* (4)*), and for any $t \geq 0$, denote by $\epsilon(t)$ the associated unbalancedness magnitude (Equation* (18)*). Then, $\epsilon(t)$ is constant through time, i.e. $\epsilon(t) = \epsilon(0)$ for all $t \geq 0$.*

Proof sketch (for complete proof see Subappendix G.12). For $l = 1, 2, \dots, L-1$, using the dynamics of $W_l(t)$ and $W_{l+1}(t)$ under gradient flow, we show that:

$$\tfrac{d}{dt}(W_l(t)W_l(t)^\top) = \tfrac{d}{dt}(W_{l+1}(t)^\top W_{l+1}(t)) \quad , \forall t \geq 0\,.$$

This implies $W_{l+1}(t)^\top W_{l+1}(t) - W_l(t)W_l(t)^\top = W_{l+1}(0)^\top W_{l+1}(0) - W_l(0)W_l(0)^\top$ for all $t \geq 0$. The proof concludes by taking nuclear norm of both sides of the latter equality, followed by maximization over $l \in \{1, 2, \dots, L-1\}$. □

Combining Lemma 3 with Lemma 1 implies that if unbalancedness magnitude is small at initialization, it remains that way throughout, and thus for every point along the optimization trajectory there exists some nearby point which is balanced (*i.e.* has unbalancedness magnitude zero). We may imagine a gradient flow trajectory emanating from such balanced point, and import certain characteristics from this imaginary trajectory to the original one. The idea of using imaginary balancedly-initialized trajectories for analyzing the unbalanced case also appears in the approach laid out in Subappendix D.1. However, whereas there only one such trajectory was employed, here there are infinitely many — one for each point in time. This allows us to maintain small distance from an imaginary trajectory (as opposed to a distance that grows exponentially with time — see Proposition 5), facilitating import of characteristics during which incurred error terms are small.

In the context of Theorems 1 and 2, the critical characteristic of trajectories originating from balanced initializations is that they do not allow the product matrix's (Equation (3)) determinant to change sign, or more specifically, its smallest singular value to cross zero. Using the aforementioned technique (proximity to imaginary balancedly-initialized trajectories), we may import an approximate version of this characteristic into trajectories whose initializations have small unbalancedness magnitude. This amounts to a bound on the rate at which the smallest singular value of the product matrix can approach zero, yielding a guaranteed time throughout which the results of Theorems 1 and 2 hold.

Theorem 3 below formalizes the logic outlined above, extending Theorem 1 to the case of unbalanced initialization (we omit here the formal extension of Theorem 2, as it is essentially the same).

Theorem 3. *Consider the setting of Theorem 1, with the assumption of balanced initialization (Equation* (5)*) removed, allowing initialization with unbalancedness magnitude $\epsilon > 0$ (Definition 3). Assume that:* (i) *the deep matrix factorization is square, i.e. its hidden dimensions are equal to 2;* (ii) *the loss at initialization is lower than that at zero, i.e. $\ell(W_{L:1}(0)) < \ell(0) = 1$, where $W_{L:1}(t)$ denotes the product matrix (Equation* (3)*) at time $t \geq 0$ of optimization; and* (iii)

$$\epsilon \leq \begin{cases} \exp\left(-\frac{2^{16}\left(\max\left\{32, \max_{l\in[L]}\|W_l(0)\|_{Fro}\right\}+1\right)^6}{(1-\sqrt{\ell_{init}})^4} \right) & \text{, \textit{if depth} } L = 2 \\ \frac{(1-\sqrt{\ell_{init}})^{128}}{2^{64L+256}L^{128}\left(\max\left\{32, \max_{l\in[L]}\|W_l(0)\|_{Fro}\right\}+1\right)^{128L-64}} & \text{, \textit{if depth} } L \geq 3 \end{cases} ,$$

where $\ell_{init} := \ell(W_{L:1}(0))$.[12] *Then, the results of Theorem 1 — bounds on (quasi-)norms, effective rank and distance from infimal rank (Equations* (8), (9) *and* (10) *respectively) — all hold at least until one of the following takes place:*

- *Optimization time t reaches:*

  $$t = \begin{cases} \frac{1}{2^{2/3}(1-\sqrt{\ell_{init}})^{4/3}} \cdot \ln\left(\frac{1}{\epsilon}\right)^{2/3} - \ln\left(\frac{e}{(1-\sqrt{\ell_{init}})\sigma_{init}}\right) & \text{, \textit{if depth} } L = 2 \\ \frac{2^{4L/3}L}{(1-\sqrt{\ell_{init}})^2} \cdot \epsilon^{-\frac{3L-8}{32L-16}} - 2^{-(5L+5)}\sigma_{init}^{-\frac{L-2}{L}} & \text{, \textit{if depth} } L \geq 3 \end{cases} , \qquad (21)$$

  where $\sigma_{init} := \min\{\sigma_{min}(W_{L:1}(0)), (1 - \sqrt{\ell_{init}})/2\}$, with $\sigma_{min}(W_{L:1}(0))$ standing for the minimal singular value of $W_{L:1}(0)$; or

[12] These assumptions are technical in nature; we defer their relaxation to future work.

- *(Quasi-)norms, effective rank and distance from infimal rank are jointly bounded as follows:*

$$\|W_{L:1}(t)\| \geq \begin{cases} \frac{\left\|\mathbf{e}_1\mathbf{e}_1^\top\right\|(1-\sqrt{\ell_{init}})^{4/3}}{2^{11}c_{\|\cdot\|}} \cdot \ln\left(\frac{1}{\epsilon}\right)^{1/3} - 12c_{\|\cdot\|}^2 \max_{i,j\in\{1,2\}} \left\|\mathbf{e}_i\mathbf{e}_j^\top\right\| & \text{, if depth } L = 2 \\ \frac{\left\|\mathbf{e}_1\mathbf{e}_1^\top\right\|(1-\sqrt{\ell_{init}})^{6/5}}{2^{4L}L^{6/5}c_{\|\cdot\|}} \cdot \epsilon^{-\frac{L}{128L-64}} - 12c_{\|\cdot\|}^2 \max_{i,j\in\{1,2\}} \left\|\mathbf{e}_i\mathbf{e}_j^\top\right\| & \text{, if depth } L \geq 3 \end{cases}, \tag{22}$$

$$\operatorname{erank}(W_{L:1}(t)) \leq \begin{cases} \inf_{W'\in\mathcal{S}} \operatorname{erank}(W') + \frac{2^9}{(1-\sqrt{\ell_{init}})^{2/3}} \cdot \ln\left(\frac{1}{\epsilon}\right)^{-1/6} & \text{, if depth } L = 2 \\ \inf_{W'\in\mathcal{S}} \operatorname{erank}(W') + \frac{2^{2L+5}L}{1-\sqrt{\ell_{init}}} \cdot \epsilon^{\frac{L}{256L-128}} & \text{, if depth } L \geq 3 \end{cases}, \tag{23}$$

$$D(W_{L:1}(t), \mathcal{M}_{\operatorname{irank}(\mathcal{S})}) \leq \begin{cases} \frac{2^{12}}{(1-\sqrt{\ell_{init}})^{4/3}} \cdot \ln\left(\frac{1}{\epsilon}\right)^{-1/3} + \sqrt{2\ell(W_{L:1}(t))} & \text{, if depth } L = 2 \\ \frac{2^{3L+4}L^{6/5}}{(1-\sqrt{\ell_{init}})^{6/5}} \cdot \epsilon^{\frac{L}{128L-64}} + \sqrt{2\ell(W_{L:1}(t))} & \text{, if depth } L \geq 3 \end{cases}, \tag{24}$$

  where $\|\cdot\|$ *is any norm or quasi-norm over matrices,* $c_{\|\cdot\|} \geq 1$ *is a constant with which* $\|\cdot\|$ *satisfies the weakened triangle inequality (see Footnote 2),* $\operatorname{erank}(\cdot)$ *stands for effective rank (Definition 1), and* $D(\cdot\,, \mathcal{M}_{\operatorname{irank}(\mathcal{S})})$ *represents distance from the infimal rank (Definition 2) of the solution set* $\mathcal{S}$ *(Equation* (7)*).*

Proof sketch (for complete proof see Subappendix G.13). By the proof of Theorem 1 (given in Subappendix G.5), its results (Equations (8), (9) and (10)) hold for any $t \geq 0$ with $\det(W_{L:1}(t)) > 0$. Bearing in mind that by assumption $\det(W_{L:1}(0)) > 0$, we let $T \in (0, \infty)$ be the initial time at which $\det(W_{L:1}(T)) = 0$ (if no such T exists, the proof concludes). Fixing an arbitrary time $\bar{t} \in [0, T]$, Lemmas 3 and 1 imply that there exists a point $(W_1', W_2', \ldots, W_L')$ which meets the balancedness condition (*i.e.* has unbalancedness magnitude zero), and is within (Frobenius) distance $\mathcal{O}(L^2 \cdot \sqrt{\epsilon})$ from $(W_1(\bar{t}), W_2(\bar{t}), \ldots, W_L(\bar{t}))$. Imagining a gradient flow path that emanates from $(W_1', W_2', \ldots, W_L')$, one may employ Lemma 5 from Subappendix G.2.1, to characterize the movement of the singular values of $W_{L:1}' := W_L' W_{L-1}' \cdots W_1'$. Continuity arguments then imply that the singular values of $W_{L:1}(\bar{t})$ move similarly (at time $\bar{t}$), allowing us to obtain an upper bound on the rate at which the minimal singular value of $W_{L:1}(\bar{t})$ can decay. Integrating this upper bound yields a lower bound on T, specified in Equation (21) as one of the possible outcomes. The continuity arguments employed require $\|(W_1(t), W_2(t), \ldots, W_L(t))\|_{Fro}$, $t \in [0, T]$, to be bounded by a certain constant. If this is not the case then necessarily at some time $t \leq T$ the unobserved entry of $W_{L:1}(t)$ is large, leading to the bounds on (quasi-)norms, effective rank and distance from infimal rank in the alternative outcome (Equations (22), (23) and (24) respectively). □

Theorem 3 states that if initialization has unbalancedness magnitude $\epsilon > 0$ (Definition 3), then the results of Theorem 1 — bounds on (quasi-)norms, effective rank and distance from infimal rank (Equations (8), (9) and (10) respectively) — are guaranteed to hold for a certain period of time (Equation (21)), or until certain terminal bounds (Equations (22), (23) and (24)) are jointly satisfied. Taking $\epsilon \to 0^+$, the aforementioned period of time tends to infinity, and the terminal bounds tend to the limits (as loss goes to zero) of the bounds in Theorem 1, meaning we effectively converge to the latter. The rate of this convergence highly depends on the depth of the matrix factorization — roughly speaking, it is proportional to a fractional power of $\ln(1/\epsilon)$ for depth 2, and to a fractional power of $1/\epsilon$ for depth 3 or more. Disregarding constants (terms that do not depend on ϵ),[13] this implies that in order to get comparable guarantees, the unbalancedness magnitude of initialization needs to be exponentially smaller with depth 2 than with depth 3 or more. We thus have a theoretical reasoning that resonates with the empirical phenomenon reported in Figure 1, by which in practical settings (gradient descent with small learning rate and near-zero initialization), the prediction of Theorem 1 — unobserved entry increasing (and therefore norms and quasi-norms increasing, with

[13] We did not attempt to optimize those; doing so is regarded as a potential direction for future work.

effective rank and distance from infimal rank decreasing) as loss decreases — sustains for much longer with depth 3 or more than it does with depth 2.[14]

E Extension to different matrix dimensions

In this appendix we outline an extension of the construction and analysis given in Subsections 3.1 and 3.2 respectively, to completion of matrices with dimensions beyond 2-by-2. The extension presented here is not unique, but rather one simple option out of many. It is demonstrated empirically in Subappendix F.1 (Figure 4).

Beginning with square matrices, for $2 \leq d \in \mathbb{N}$, consider completion of a d-by-d matrix based on the following observations:

$$\begin{aligned} \Omega &= \{1,\ldots,d\} \times \{1,\ldots,d\} \setminus \{(1,1)\}\,, \\ b_{i,j} &= \begin{cases} 1 & \text{, if } i = j \geq 3 \text{ or } (i,j) \in \{(1,2),(2,1)\} \\ 0 & \text{, otherwise} \end{cases} \quad \text{, for } (i,j) \in \Omega\,, \end{aligned} \tag{25}$$

where, as in Section 2, Ω represents the set of observed locations, and $\{b_{i,j} \in \mathbb{R}\}_{(i,j)\in\Omega}$ the corresponding set of observed values. The solution set for this problem (*i.e.* the set of matrices zeroing the loss in Equation (1)) is:

$$\mathcal{S}_d := \left\{ \begin{pmatrix} w_{1,1} & 1 & 0 & 0 \cdots & 0 \\ 1 & 0 & 0 & 0 \cdots & 0 \\ 0 & 0 & 1 & 0 \cdots & 0 \\ 0 & 0 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & & \ddots \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \in \mathbb{R}^{d,d} : w_{1,1} \in \mathbb{R} \right\}. \tag{26}$$

Observing $\mathcal{S}_d$, while comparing to the solution set $\mathcal{S}$ in our original construction (Equation (7)), we see that the former has a 2-by-2 block diagonal structure, with the top-left block holding the latter, and the bottom-right block set to identity. This implies that $d-2$ of the singular values along $\mathcal{S}_d$ are fixed to one, and the remaining two are identical to the singular values along $\mathcal{S}$. Results analogous to Propositions 1 and 2 can therefore easily be proven. Since the determinant along $\mathcal{S}_d$ is bounded below and away from zero (it is equal to -1), approaching $\mathcal{S}_d$ while having positive determinant necessarily means that absolute value of unobserved entry (*i.e.* of the entry in location $(1,1)$) grows towards infinity. Combining this with the fact that the product matrix (Equation (3)) of a depth $L \geq 2$ matrix factorization maintains the sign of its determinant (see Lemma 6 in Subappendix G.2.1), results analogous to Theorem 1 and Corollary 1 may readily be established. That is, one may show that, with probability 0.5 or more over random near-zero initialization, gradient descent with small learning rate drives *all* norms (and quasi-norms) *towards infinity*, while essentially driving rank towards its minimum.

Moving on to the rectangular case, for $2 \leq d, d' \in \mathbb{N}$, consider completion of a d-by-d' matrix based on the

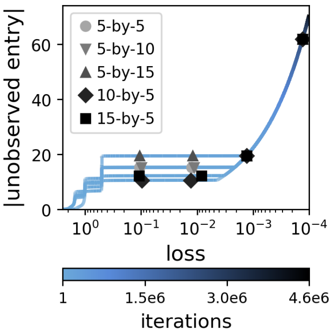

Figure 4: Phenomenon of implicit regularization in matrix factorization driving *all* norms (and quasi-norms) *towards infinity* extends to arbitrary matrix dimensions. Appendix E outlines an extension of the construction and analysis given in Subsections 3.1 and 3.2 respectively, to completion of matrices with arbitrary dimensions. The extension implies that for any $2 \leq d, d' \in \mathbb{N}$, when applying matrix factorization to the specified d-by-d' matrix completion problem, decreasing loss, *i.e.* fitting observations, can lead absolute value of unobserved entry to increase (which in turn means norms and quasi-norms increase). This is demonstrated in the plot above, which for representative runs corresponding to different choices of d and d', shows absolute value of unobserved entry as a function of the loss (Equation 1), with iteration number encoded by color. Runs were obtained with a depth 3 matrix factorization initialized randomly by an unbalanced (layer-wise independent) distribution, with the latter's standard deviation and the learning rate for gradient descent set to the smallest values used for depth 3 in Figure 1 (other settings we evaluated produced similar results). For further implementation details see Subappendix F.2.1.

[14] Note that this phenomenon takes place even when initializations are perfectly balanced (*i.e.* have unbalancedness magnitude zero), the reason being the discrepancy between gradient descent with small learning rate and gradient flow. Specifically, while the latter would have conserved the balancedness throughout (Lemma 3), the former will (generically) lead to positive unbalancedness magnitude immediately after its commencement.

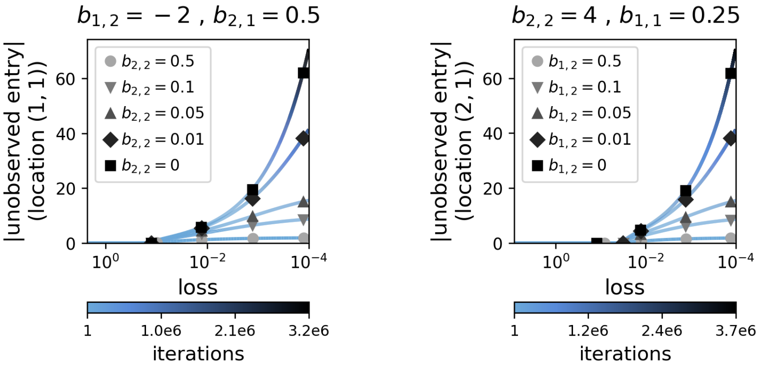

Figure 5: Phenomenon of implicit regularization in matrix factorization driving *all* norms (and quasi-norms) *towards infinity* is robust to perturbations. Our analysis (Subsection 3.4) implies that, when applying matrix factorization to the matrix completion problem defined in Subsection 3.1, even if observations are perturbed and repositioned, decreasing loss, *i.e.* fitting them, leads absolute value of unobserved entry to increase (which in turn means norms and quasi-norms increase). Specifically, with $(i, j) \in \{1, 2\} \times \{1, 2\}$ representing the unobserved location and $\bar{i} := 3 - i$, $\bar{j} := 3 - j$, Theorem 2 implies that: *(i)* if the diagonally-opposite observation $b_{\bar{i},\bar{j}}$ is unperturbed (stays at zero), the adjacent ones $b_{i,\bar{j}}, b_{\bar{i},j}$ can take on *any* non-zero values, and as long as at initialization the sign of the product matrix's (Equation 3) determinant accords with that of $b_{i,\bar{j}} \cdot b_{\bar{i},j}$, the absolute value of unobserved entry will grow to infinity; and *(ii)* the extent to which absolute value of unobserved entry grows gracefully recedes as $b_{\bar{i},\bar{j}}$ is perturbed away from zero. This is demonstrated in the plots above, which for representative runs, show absolute value of unobserved entry as a function of the loss (Equation 1), with iteration number encoded by color. Each plot corresponds to a different choice of (i, j) and a different assignment for $b_{i,\bar{j}}, b_{\bar{i},j}$, presenting runs with varying values for $b_{\bar{i},\bar{j}}$. Runs were obtained with a depth 3 matrix factorization initialized randomly by an unbalanced (layer-wise independent) distribution, with the latter's standard deviation and the learning rate for gradient descent set to the smallest values used for depth 3 in Figure 1 (other settings we evaluated produced similar results). For further implementation details see Subappendix F.2.1.

same observations as in Equation (25), but with additional zero observations such that only the entry in location $(1, 1)$ is unobserved. The singular values along the solution set for this problem are the same as those along $\mathcal{S}_d$ (Equation (26)). Moreover, assuming without loss of generality that $d \leq d'$, if a matrix factorization applied to this problem is initialized such that its product matrix holds zeros in columns $d + 1$ to d', then a dynamical characterization from [6] (restated as Lemma 4 in Subappendix G.2.1), along with the structure of the loss (Equation (1)), ensure the leftmost d-by-d submatrix of the product matrix evolves precisely as in the square case discussed above, while the remaining columns ($d + 1$ to d') stay at zero. Results thus carry over from the square to the rectangular case.

F Further experiments and implementation details

F.1 Further experiments

Figures 4 and 5 supplement Figure 1 from Subsection 4.1, by demonstrating empirically that the phenomenon of implicit regularization in matrix factorization driving all norms (and quasi-norms) towards infinity is, respectively: *(i)* applicable to arbitrary matrix dimensions, as outlined in Appendix E; and *(ii)* robust to perturbations, as proven in Subsection 3.4. Figure 6 supplements Figure 2 from Subsection 4.2, further demonstrating that gradient descent over tensor factorization exhibits an implicit regularization towards low (tensor) rank.

F.2 Implementation details

Below we provide a full description of implementation details omitted from our experimental reports (Section 4 and Subappendix F.1). Source code for reproducing our results and figures, based on the PyTorch framework ([68]), can be found at `https://github.com/noamrazin/imp_reg_dl_not_norms`.

F.2.1 Deep matrix factorization (Figures 1, 4, and 5)

In all experiments with deep matrix factorization, hidden dimensions were set to the minimal value ensuring unconstrained search space, *i.e.* to the minimum between the number of rows and the number of columns in the matrix to complete. Gradient descent was run with fixed learning rate until loss (Equation (1)) reached a value lower than 10^{-4} or $5 \cdot 10^6$ iterations elapsed. Both balanced (Equa-

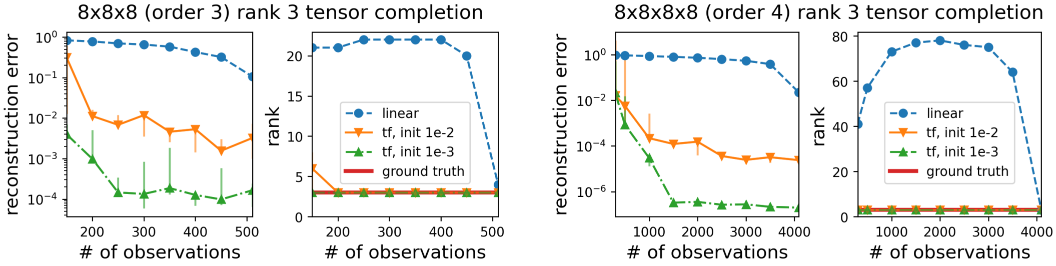

Figure 6: Gradient descent over tensor factorization exhibits an implicit regularization towards low (tensor) rank. This figure is identical to Figure 2, except that the experiments it portrays had ground truth tensors of rank 3 (instead of 1). For further details see caption of Figure 2, as well as Subappendix F.2.2.

tion (5)) and unbalanced (layer-wise independent) random initializations were calibrated according to a desired standard deviation $\alpha > 0$ for the entries of the initial product matrix (Equation (3)). Namely: *(i)* under unbalanced initialization, entries of all weight matrices were sampled independently from a Gaussian distribution with zero mean and standard deviation $(\alpha^2/\bar{d}^{L-1})^{1/2L}$, where L stands for the depth of the factorization, and $\bar{d}$ for the size of its hidden dimensions; and *(ii)* under balanced initialization, we used Procedure 1 from [7], based on a Gaussian distribution with independent entries, zero mean and standard deviation α. In accordance with the description in Appendix E, if the matrix to complete was rectangular, we ensured that excess rows or columns of the initial product matrix held zeros, by clearing (setting to zero) corresponding rows or columns of the initial leftmost or rightmost (respectively) matrix in the factorization.[15] Random initializations were repeated until the determinant of the initial product matrix (or of its top-left $\min\{d, d'\}$-by-$\min\{d, d'\}$ submatrix if its size was d-by-d' with $d \neq d'$) was of the necessary sign,[16] taking two attempts on average. In the experiment reported by Figure 1, runs with matrix factorization depths 2 and 3 were carried out with learning rates $\{6 \cdot 10^{-2}, 3 \cdot 10^{-2}, 9 \cdot 10^{-3}, 6 \cdot 10^{-3}, 3 \cdot 10^{-3}, 9 \cdot 10^{-4}\}$ and corresponding standard deviations for initialization $\{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}\}$. Factorizations of depth 4 were slightly more sensitive to changes in learning rate, thus we refined attempted values to $\{6 \cdot 10^{-3}, 4.5 \cdot 10^{-3}, 3 \cdot 10^{-3}, 1.5 \cdot 10^{-3}, 10^{-3}\}$, with corresponding standard deviations for initialization $\{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$.

F.2.2 Tensor factorization (Figures 2 and 6)

In all experiments with tensor factorization (Equation (12)), the number of terms R was set to ensure an unconstrained search space, *i.e.* it was set to 8^2 and 8^3 for tensor sizes 8-by-8-by-8 and 8-by-8-by-8-by-8 respectively.[17] Horizontal axes in all plots begin from the smallest number of observations producing stable results, and end when all entries but one are observed. Specifically: *(i)* in the experiments with rank 1 ground truth tensors (Figure 2), the number of observations ranged over $\{50, 100, 150, \ldots, 400, 450, 511\}$ and $\{100, 500, 1000, 1500, \ldots, 3000, 3500, 4095\}$ for orders 3 and 4 respectively; and *(ii)* for experiments with rank 3 ground truth tensors (Figure 6), the minimal number of observations was increased threefold (*i.e.* ranges of $\{150, 200, 250, \ldots, 400, 450, 511\}$ and $\{300, 500, 1000, 1500, \ldots, 3000, 3500, 4095\}$ were used for orders 3 and 4 respectively). Gradient descent was run until the mean squared error over observations reached a value lower than 10^{-6} or 10^6 iterations elapsed. For initialization, weights were sampled independently from a Gaussian distribution with zero mean and varying standard deviation. In particular, five trials (differing in random seed) were conducted for each standard deviation in the range $\{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$. To facilitate more efficient experimentation, we employed an adaptive learning rate scheme, where at each iteration a base learning rate of 10^{-2} was divided by the square root of an exponential moving average of squared gradient norms. That is, with base learning rate $\eta = 10^{-2}$ and weighted average coefficient $\beta = 0.99$, at iteration t the learning rate was set to $\eta_t = \eta / (\sqrt{\gamma_t / (1 - \beta^t)} + 10^{-6})$, where $\gamma_t = \beta \cdot \gamma_{t-1} + (1 - \beta) \cdot \sum_{r=1}^{R} \sum_{n=1}^{N} \| \partial / \partial \mathbf{w}_r^{(n)} \ell(\{\mathbf{w}_r^{(n)}(t)\}_{r,n}) \|_F^2$, with $\gamma_0 = 0$ and $\ell(\cdot)$ standing for the mean squared error over observations. We emphasize that only the learning rate (step size) is affected by this scheme, not the direction of movement. Comparisons between the scheme and a fixed

[15] That is, if the matrix to complete had size d-by-d' with $d \neq d'$, we cleared rows $d' + 1$ to d of $W_L(0)$ if $d > d'$, and columns $d + 1$ to d' of $W_1(0)$ if $d' > d$.

[16] Positive for the experiments reported by Figures 1 and 4, and negative for those reported by Figure 5.

[17] As shown in [37], for any $d_1, d_2, \ldots, d_N \in \mathbb{N}$, using $R = (\Pi_{n=1}^{N} d_n) / \max\{d_n\}_{n=1}^{N}$ suffices for expressing all tensors in $\mathbb{R}^{d_1, d_2, \ldots, d_N}$.

(small) learning rate schedule have shown no noticeable impact on the end result, with significant difference in terms of run time.

When referring to tensor rank, we mean the classic *CP-rank* (see [51]). While exact inference of this rank is in the worst case computationally hard (*cf.* [39]), in practice, a standard way to estimate it is by the minimal number of terms (R in Equation (12)) for which the Alternating Least Squares (ALS) algorithm achieves reconstruction (mean squared) error below a certain threshold (see [51] for further details). We follow this method with a threshold of 10^{-6}. Generating a ground truth rank R^* tensor $\mathcal{W}^* \in \mathbb{R}^{d_1, d_2, \ldots, d_N}$ was done by computing:

$$\mathcal{W}^* = \sum\nolimits_{r=1}^{R^*} \mathbf{w}_r^{*(1)} \otimes \mathbf{w}_r^{*(2)} \otimes \cdots \otimes \mathbf{w}_r^{*(N)} \quad , \, \mathbf{w}_r^{*(n)} \in \mathbb{R}^{d_n} \, , \, r = 1, 2, ..., R^* \, , \, n = 1, 2, ..., N \, ,$$

with $\{\mathbf{w}_r^{*(n)}\}_{r=1}^{R^*}{}_{n=1}^{N}$ drawn independently from the standard normal distribution. After every such generation, we estimated the rank of the obtained tensor (its construction only ensures a rank of *at most* R^*), and repeated the process if it was smaller than R^*. For convenience, we subsequently normalized the ground truth tensor to be of unit Frobenius norm.

G Deferred proofs

G.1 Notations

We define a few notational conventions that will be used throughout our proofs. For $N \in \mathbb{N}$, let $[N]$ denote the set $\{1, 2, \ldots, N\}$. Let $\{\mathbf{e}_i\}_{i=1}^{d} \subset \mathbb{R}^d$ be the standard basis vectors, *i.e.* $\mathbf{e}_i$ holds 1 in its i'th coordinate and 0 elsewhere. The singular values of a matrix $W \in \mathbb{R}^{d,d'}$ are denoted by $\sigma_1(W) \geq \ldots \geq \sigma_{\min\{d,d'\}}(W) \geq 0$, where by convention $\sigma_i(W) := 0$ for $i > \min\{d, d'\}$. Similarly, the eigenvalues of a symmetric matrix $W \in \mathbb{R}^{d,d}$ are denoted by $\lambda_1(W) \geq \ldots \geq \lambda_d(W)$. We let $\|W\|_{S_p}$, with $p \in (0, \infty]$, stand for the Schatten-p (quasi-)norm of a matrix $W \in \mathbb{R}^{d,d'}$, and denote by $\|W\|_F$ the special case $p = 2$, *i.e.* the Frobenius norm. The Euclidean norm of a vector $w \in \mathbb{R}^d$ is denoted by $\|w\|_2$. Since norms are a special case of quasi-norms, when providing results applicable to both, only the latter is explicitly treated. To admit a compact representation of matrix products, given $1 \leq a \leq b \leq L$ and matrices $W_1, W_2, \ldots, W_L$ for which the product $W_L W_{L-1} \cdots W_1$ is defined, we denote:

$$\prod\nolimits_{a}^{r=b} W_r := W_b \cdots W_a \, ,$$
$$\prod\nolimits_{r=a}^{b} W_r^\top := W_a^\top \cdots W_b^\top \, .$$

By definition, if $a > b$, then both $\prod_a^{r=b} W_r$ and $\prod_{r=a}^{b} W_r^\top$ are identity matrices, with size to be inferred by context. The k'th derivative of a function (from $\mathbb{R}$ to $\mathbb{R}$) $f(t)$ is denoted by $f^{(k)}(t)$, with $f^{(0)}(t) := f(t)$ by convention. For consistency with differential equations literature, when the variable t is regarded as a time index, we also denote the first order derivative by $\dot{f}(t)$. Lastly, when clear from context, a time index t will often be omitted.

G.2 Useful lemmas

G.2.1 Deep matrix factorization

For completeness, we include the following result from [6], which characterizes the evolution of the product matrix under gradient flow on a deep matrix factorization:

Lemma 4 (adaptation of Theorem 1 in [6]). *Let $\ell : \mathbb{R}^{d,d'} \to \mathbb{R}_{\geq 0}$ be an analytic*[18] *loss, overparameterized by a depth L matrix factorization:*

$$\phi(W_1, \ldots, W_L) = \ell(W_L W_{L-1} \cdots W_1) \, .$$

Suppose we run gradient flow over the factorization:

$$\dot{W}_l(t) := \tfrac{d}{dt} W_l(t) = -\tfrac{\partial}{\partial W_l} \phi(W_1(t), \ldots, W_L(t)) \quad , \, t \geq 0 \, , \, l = 1, \ldots, L \, ,$$

[18] An infinitely differentiable function $f : \mathcal{D} \to \mathbb{R}$ is *analytic* if at every $x \in \mathcal{D}$ its Taylor series converges to it on some neighborhood of x (see [52] for further details). Specifically, the matrix completion loss considered (Equation (1)) is analytic.

with a balanced initialization, i.e.:

$$W_{l+1}(0)^\top W_{l+1}(0) = W_l(0) W_l(0)^\top \quad , \; l = 1, \ldots, L-1 \, .$$

Then, the product matrix $W_{L:1}(t) = W_L(t) \cdots W_1(t)$ *obeys the following dynamics:*

$$\dot{W}_{L:1}(t) = -\sum\nolimits_{l=1}^{L} \left[W_{L:1}(t) W_{L:1}(t)^\top \right]^{\frac{l-1}{L}} \cdot \nabla \ell\big(W_{L:1}(t)\big) \cdot \left[W_{L:1}(t)^\top W_{L:1}(t) \right]^{\frac{L-l}{L}} \; ,$$

where $[\,\cdot\,]^\beta$, $\beta \in \mathbb{R}_{\geq 0}$, *stands for a power operator defined over positive semidefinite matrices (with* $\beta = 0$ *yielding identity by definition).*[19]

Additionally, recall from [8] the following characterization for the singular values of $W_{L:1}(t)$:

Lemma 5 (adaptation of Lemma 1 and Theorem 3 in [8])**.** *Consider the setting of Lemma 4 for depth* $2 \leq L \in \mathbb{N}$. *Then, there exist analytical functions* $\{\sigma_r : [0, \infty) \to \mathbb{R}\}_{r=1}^{\min\{d,d'\}}$, $\{\mathbf{u}_r : [0, \infty) \to \mathbb{R}^d\}_{r=1}^{\min\{d,d'\}}$ *and* $\{\mathbf{v}_r : [0, \infty) \to \mathbb{R}^{d'}\}_{r=1}^{\min\{d,d'\}}$ *such that:*

$$\sigma_r(t) \geq 0 \;\; , \;\; \mathbf{u}_r(t)^\top \mathbf{u}_{r'}(t) = \mathbf{v}_r(t)^\top \mathbf{v}_{r'}(t) = \begin{cases} 1 & , r = r' \\ 0 & , r \neq r' \end{cases} \quad , \; t \geq 0 \; , \; r, r' \in [\min\{d, d'\}]$$

$$W_{L:1}(t) = \sum\nolimits_{r=1}^{\min\{d,d'\}} \sigma_r(t) \mathbf{u}_r(t) \mathbf{v}_r(t)^\top \qquad ,$$

i.e. $\sigma_r(t)$ *are the singular values of* $W_{L:1}(t)$*, and* $\mathbf{u}_r(t), \mathbf{v}_r(t)$ *are corresponding left and right (respectively) singular vectors. Furthermore, the singular values* $\sigma_r(t)$ *evolve by:*

$$\dot{\sigma}_r(t) = -L \cdot \big(\sigma_r^2(t)\big)^{1-1/L} \cdot \left\langle \nabla \ell \left(W_{L:1}(t) \right), \mathbf{u}_r(t) \mathbf{v}_r(t)^\top \right\rangle \quad , \; r = 1, \ldots, \min\{d, d'\} \, . \qquad (27)$$

We rely on this result to establish that for square product matrices the sign of $\det(W_{L:1}(t))$ does not change throughout time.

Lemma 6. *Consider the setting of Lemma 4 with depth* $2 \leq L \in \mathbb{N}$ *and* $d = d'$. *Then, the determinant of* $W_{L:1}(t)$ *has the same sign as its initial value* $\det(W_{L:1}(0))$. *That is,* $\det(W_{L:1}(t))$ *is identically zero if* $\det(W_{L:1}(0)) = 0$*, is positive if* $\det(W_{L:1}(0)) > 0$*, and is negative if* $\det(W_{L:1}(0)) < 0$*.*

Proof. We prove an analogous claim for the singular values of $W_{L:1}(t)$, from which the lemma readily follows. That is, for $r \in [d]$, the singular value $\sigma_r(t)$ is identically zero if $\sigma_r(0) = 0$, and is positive if $\sigma_r(0) > 0$.

For conciseness, define $g(t) := -L \cdot \left\langle \nabla \ell \left(W_{L:1}(t) \right), \mathbf{u}_r(t) \mathbf{v}_r(t)^\top \right\rangle$. Invoking Lemma 5, let us solve the differential equation for $\sigma_r(t)$. If $L = 2$, the solution to Equation (27) is $\sigma_r(t) = \sigma_r(0) \cdot \exp\left(\int_{t'=0}^{t} g(t') dt' \right)$. Clearly, $\sigma_r(t)$ is either identically zero or positive according to its initial value. If $L > 2$, Equation (27) is solved by:

$$\sigma_r(t) = \begin{cases} \left(\sigma_r(0)^{\frac{2}{L}-1} + \left(\frac{2}{L} - 1\right) \int_{t'=0}^{t} g(t') dt' \right)^{\frac{1}{\frac{2}{L}-1}} & , \sigma_r(0) > 0 \\ 0 & , \sigma_r(0) = 0 \end{cases} .$$

As before, if $\sigma_r(0) = 0$, then $\sigma_r(t) = 0$ for all $t \geq 0$. If $\sigma_r(0) > 0$, divergence in finite time of $\sigma_r(t)$ is possible, however, its positivity is preserved until that occurs nonetheless.

Turning our attention to the determinant of $W_{L:1}(t)$, suppose $\det(W_{L:1}(0)) = 0$. Then, $W_{L:1}(0)$ has a singular value which is 0, and for all t that singular value and the determinant remain 0. If $\det(W_{L:1}(0)) \neq 0$, the product matrix remains full rank for all t. The proof then immediately follows from the continuity of $\det(W_{L:1}(t))$. □

We will also make use of the following lemmas:

[19] As discussed in Section 2, mounting empirical and theoretical evidence suggest a close match between the predictions of gradient flow with balanced initialization, and its practical realization via gradient descent with small learning rate and near-zero initialization (*cf.* [6, 7, 45]). It was recently argued in [70] that certain aspects of balancedness do not transfer from gradient flow to gradient descent. However, the definitions in [70] deviate from the conventional ones, hence its conclusions are not applicable to standard settings.

Lemma 7 (adapted from [8]). *Under the setting of Lemma 4,* $W_1(t), W_2(t), \ldots, W_L(t), W_{L:1}(t)$ *and* $\nabla\ell(W_{L:1}(t))$ *are analytic functions of* t*.*

Proof. Analytic functions are closed under summation, multiplication, and composition. The analyticity of $\ell(\cdot)$ therefore implies that $\phi(\cdot)$ (Equation (2)) is analytic as well. From Theorem 1.1 in [41], it then follows that under gradient flow (Equation (4)) $W_1(t), W_2(t), \ldots, W_L(t)$ are analytic functions of t. Lastly, the aforementioned closure properties imply that $W_{L:1}(t)$ and $\nabla\ell(W_{L:1}(t))$ are also analytic in t. □

Lemma 8. *For any matrices* $\{W_l \in \mathbb{R}^{d_l, d_{l-1}}\}_{l=1}^L$*, with* $d_0, d_1, \ldots, d_L \in \mathbb{N}$*, that are balanced, i.e. that meet* $W_{l+1}^\top W_{l+1} = W_l W_l^\top$ *for all* $l \in [L-1]$*, it holds that* $\sigma_i(W_{L:1}) = \sigma_i(W_l)^L$ *for all* $l \in [L]$ *and* $i \in [\min\{d_L, d_0\}]$*.*

Proof. We construct singular value decompositions for $W_1, W_2, \ldots, W_L$ in an iterative process as follows. First, let $W_1 = U_1 \Sigma_1 V_1^\top$ be an arbitrary singular value decomposition of W_1, *i.e.* $U_1 \in \mathbb{R}^{d_1,d_1}, V_1 \in \mathbb{R}^{d_0,d_0}$ are orthogonal matrices, and $\Sigma_1 \in \mathbb{R}_{\geq 0}^{d_1,d_0}$ is rectangular-diagonal holding the singular values of W_1 in non-increasing order. Then, for $l = 2, 3, \ldots, L$, balancedness of $W_1, W_2, \ldots, W_L$ and Lemma 13 imply that W_l has a singular value decomposition $U_l \Sigma_l U_{l-1}^\top$, where: $U_l \in \mathbb{R}^{d_l d_l}$ is orthogonal; $\Sigma_l \in \mathbb{R}_{\geq 0}^{d_l, d_{l-1}}$ is rectangular-diagonal holding the singular values of W_l in non-increasing order; and $\sigma_i(W_l) = (\Sigma_l)_{i,i} = (\Sigma_{l-1})_{i,i} = \sigma_i(W_{l-1})$ for $i \in [\min\{d_l, d_{l-1}, d_{l-2}\}]$, with the remaining diagonal entries of Σ_l and Σ_{l-1} being 0. With $\{U_l\}_{l=1}^L, \{\Sigma_l\}_{l=1}^L, V_1$ as described above, the product matrix can be written as:

$$W_{L:1} = \prod\nolimits_1^{l=L} W_l = \left[\prod\nolimits_2^{l=L} U_l \Sigma_l U_{l-1}^\top\right] \cdot U_1 \Sigma_1 V_1^\top = U_L \cdot \prod\nolimits_1^{l=L} \Sigma_l \cdot V_1^\top .$$

That is, $U_L \cdot \prod_1^{l=L} \Sigma_l \cdot V_1^\top$ is a singular value decomposition of $W_{L:1}$, and $\sigma_i(W_{L:1}) = (\prod_{l=1}^L \Sigma_l)_{i,i}$ for all $i \in [\min\{d_L, d_0\}]$.

Fix $l \in [L]$ and $i \in [\min\{d_L, d_0\}]$. If $i > \min\{d_{l'}\}_{l'=0}^L$, for any $l' \in [L]$ with $\min\{d_{l'}, d_{l'-1}\} \geq i$, by our construction it holds that $(\Sigma_{l'})_{i,i} = 0$. Hence, recalling that by convention $\sigma_i(W_l) = 0$ if $i > \min\{d_l, d_{l-1}\}$, we may conclude that $\sigma_i(W_{L:1}) = \sigma_i(W_l)^L = 0$. Otherwise, if $i \leq \min\{d_{l'}\}_{l'=0}^L$, the fact that $\sigma_i(W_{l'}) = (\Sigma_{l'})_{i,i} = (\Sigma_{l'-1})_{i,i} = \sigma_i(W_{l'-1})$ for all $l' = 2, 3, \ldots, L$ implies that $\sigma_i(W_{L:1}) = \sigma_i(W_l)^L$. □

G.2.2 Technical

Included below are a few technical lemmas used in our analyses.

Lemma 9. *Let* $h : [0,1] \to \mathbb{R}$ *be the binary entropy function* $h(p) := -p \cdot \ln(p) - (1-p)\ln(1-p)$*, where by convention* $0 \cdot \ln(0) = 0$*. Then, for all* $p \in [0,1]$*:*

$$h(p) \leq 2\sqrt{p} .$$

Proof. We present a tighter inequality, $h(p) \leq 2\sqrt{p(1-p)}$, from which the proof immediately follows since $2\sqrt{p(1-p)} \leq 2\sqrt{p}$ for $p \in [0,1]$.

Define the function $f(p) := \frac{h(p)^2}{p(1-p)}$ over the open interval $(0,1)$. Differentiating it with respect to p we have:

$$\frac{d}{dp} f(p) = \frac{(-p \cdot \ln(p))^2 - (-(1-p) \cdot \ln(1-p))^2}{p^2(1-p)^2} .$$

Introducing $g(p) := -p \cdot \ln(p)$, we show that $g(p)^2 > g(1-p)^2$ for all $p \in (0, \frac{1}{2})$. It is easily verified that $g(p) - g(1-p)$ is concave on the interval $(0, \frac{1}{2})$ (second derivative is negative). Since for $p = 0$ and $p = 1/2$ we have exactly $g(p) - g(1-p) = 0$, it holds that $g(p) - g(1-p) \geq 0$ and $g(p)^2 \geq g(1-p)^2$ for all $p \in (0, \frac{1}{2})$. Noticing $\frac{d}{dp} f(p) = \left(g(p)^2 - g(1-p)^2\right) / p^2(1-p)^2$, it follows that $f(\cdot)$ is monotonically non-decreasing on $(0, \frac{1}{2})$. Due to the fact that $f(p) = f(1-p)$, it is non-increasing on $(\frac{1}{2}, 1)$, and attains its maximal value over $(0,1)$ at $p = \frac{1}{2}$. Putting it all together, for $p \in (0,1)$ we have:

$$h(p) \leq \sqrt{p(1-p)} \cdot \sqrt{f(1/2)} = 2\ln(2) \cdot \sqrt{p(1-p)} \leq 2\sqrt{p(1-p)} ,$$

and for $p = 0, 1$ there is exact equality, completing the proof. $\square$

Lemma 10. *Let $f, g : [0, \infty) \to \mathbb{R}$ be real analytic functions (see Footnote 18) such that $f^{(k)}(0) = g^{(k)}(0)$ for all $k \in \mathbb{N} \cup \{0\}$. Then, $f(t) = g(t)$ for all $t \geq 0$.*

Proof. Define the function $h(t) := f(t) - g(t)$. Since analytic functions are closed under subtraction, $h(\cdot)$ is analytic as well. An analytic function with all zero derivatives at a point is constant on the corresponding connected component. Noticing that $h^{(k)}(0) = 0$ for all $k \in \mathbb{N} \cup \{0\}$, we may conclude that $h(t) = 0$ and $f(t) = g(t)$ for all $t \geq 0$. $\square$

Lemma 11. *Let $A, B \in \mathbb{R}^{d,d}$, and suppose B is positive semidefinite. Then,*

$$\mathrm{Tr}(A^\top BA) \geq \lambda_1(B) \cdot \sigma_d(A)^2 \,.$$

Proof. The matrix $A^\top BA$ is positive semidefinite since for all $\mathbf{y} \in \mathbb{R}^d$ we have:

$$\mathbf{y}^\top A^\top BA\mathbf{y} = (A\mathbf{y})^\top B(A\mathbf{y}) \geq 0 \,.$$

Therefore, $\mathrm{Tr}(A^\top BA) \geq \lambda_1(A^\top BA)$. Let $B = ODO^\top$ be an orthogonal eigenvalue decomposition of B, *i.e.* $O \in \mathbb{R}^{d,d}$ is an orthogonal matrix with columns $\{\mathbf{o}_i\}_{i=1}^d$ and $D \in \mathbb{R}^{d,d}$ is diagonal holding the non-negative eigenvalues of B. Additionally, let $A = U\Sigma V^\top$ be a singular value decomposition of A, where $U, V \in \mathbb{R}^{d,d}$ are orthogonal matrices, and $\Sigma \in \mathbb{R}^{d,d}_{\geq 0}$ is diagonal holding the singular values of A. For any unit vector (with respect to the ℓ_2 norm) $\mathbf{y} \in \mathbb{R}^d$ it holds that:

$$\mathbf{y}^\top A^\top BA\mathbf{y} = \sum_{i=1}^{d} \lambda_i(B)(\mathbf{o}_i^\top A\mathbf{y})^2 \geq \lambda_1(B)(\mathbf{o}_1^\top A\mathbf{y})^2 \,.$$

Replacing A with its singular value decomposition and choosing $\mathbf{y} = VU^\top \mathbf{o}_1$:

$$\lambda_1(B)(\mathbf{o}_1^\top A\mathbf{y})^2 = \lambda_1(B)(\mathbf{o}_1^\top U\Sigma U^\top \mathbf{o}_1)^2 \,.$$

Recalling that for any unit vector the quadratic form of a symmetric matrix is bounded by the maximal and minimal eigenvalues completes the proof:

$$\mathrm{Tr}(A^\top BA) \geq \lambda_1(A^\top BA) \geq \lambda_1(B)(\mathbf{o}_1^\top U\Sigma U^\top \mathbf{o}_1)^2 \geq \lambda_1(B) \cdot \sigma_d(A)^2 \,.$$

$\square$

Lemma 12. *For any $\{A_l \in \mathbb{R}^{d_l, d_{l-1}}\}_{l=1}^L$ and $\{B_l \in \mathbb{R}^{d_l, d_{l-1}}\}_{l=1}^L$, where $d_0, d_1, \ldots, d_L \in \mathbb{N}$, it holds that:*

$$\left\| \prod_{1}^{l=L} A_l - \prod_{1}^{l=L} B_l \right\|_F \leq \sum_{l=1}^{L} \|A_l - B_l\|_F \cdot \prod_{r \neq l} \max\{\|A_r\|_F \,, \|B_r\|_F\} \,.$$

Proof. The proof is by induction over $L \in \mathbb{N}$. For $L = 1$, the claim is trivial. Assuming it holds for $L - 1 \geq 1$, we prove that it holds for L as well:

$$\begin{aligned}
\left\| \prod_{1}^{l=L} A_l - \prod_{1}^{l=L} B_l \right\|_F &= \left\| \prod_{1}^{l=L} A_l - B_L \cdot \prod_{1}^{l=L-1} A_l + B_L \cdot \prod_{1}^{l=L-1} A_l - \prod_{1}^{l=L} B_l \right\|_F \\
&\leq \|A_L - B_L\|_F \cdot \left\| \prod_{1}^{l=L-1} A_l \right\|_F + \|B_L\|_F \cdot \left\| \prod_{1}^{l=L-1} A_l - \prod_{1}^{l=L-1} B_l \right\|_F \\
&\leq \|A_L - B_L\|_F \cdot \prod_{r \neq L} \max\{\|A_r\|_F \,, \|B_r\|_F\} \\
&\quad + \max\{\|A_L\|_F \,, \|B_L\|_F\} \cdot \left\| \prod_{1}^{l=L-1} A_l - \prod_{1}^{l=L-1} B_l \right\|_F \,,
\end{aligned}$$

The proof concludes by applying the inductive assumption for $L - 1$. $\square$

Lemma 13. *For $B \in \mathbb{R}^{d,k}, A \in \mathbb{R}^{k,d'}$ such that $B^\top B = AA^\top$, let $U_A \Sigma_A V_A^\top$ be a singular value decomposition of A, i.e. $U_A \in \mathbb{R}^{k,k}, V_A \in \mathbb{R}^{d',d'}$ are orthogonal matrices, and $\Sigma_A \in \mathbb{R}_{\geq 0}^{k,d'}$ is rectangular-diagonal holding the singular values of A in non-increasing order. Then, there exist an orthogonal $U_B \in \mathbb{R}^{d,d}$ and a rectangular-diagonal $\Sigma_B \in \mathbb{R}_{\geq 0}^{d,k}$ such that:*

- *$(\Sigma_B)_{i,i} = (\Sigma_A)_{i,i}$ for any $i \in [\min\{d, k, d'\}]$, with the remaining diagonal entries of Σ_A and Σ_B being 0; and*
- *$B = U_B \Sigma_B U_A^\top$, i.e. $U_B \Sigma_B U_A^\top$ is a singular value decomposition of B.*

Proof. For any matrix W it holds that $\mathrm{rank}(W) = \mathrm{rank}(WW^\top) = \mathrm{rank}(W^\top W)$. Therefore, the fact that $B^\top B = AA^\top$ implies $\mathrm{rank}(A) = \mathrm{rank}(B)$. Denote $r := \mathrm{rank}(A)$. For $i \in [k]$, let $\mathbf{u}_A^{(i)} \in \mathbb{R}^k$ be the i'th column vector of U_A. Furthermore, for $i \in [r]$ define $\mathbf{u}_B^{(i)} := \frac{1}{(\Sigma_A)_{i,i}} B \mathbf{u}_A^{(i)} \in \mathbb{R}^d$. We claim that $\mathbf{u}_B^{(1)}, \mathbf{u}_B^{(2)}, \ldots, \mathbf{u}_B^{(r)}$ form an orthonormal set. Indeed, for any $i, j \in [r]$:

$$\left\langle \mathbf{u}_B^{(i)}, \mathbf{u}_B^{(j)} \right\rangle = \frac{1}{(\Sigma_A)_{i,i}(\Sigma_A)_{j,j}} \left(\mathbf{u}_A^{(i)}\right)^\top B^\top B \mathbf{u}_A^{(j)} = \frac{(\Sigma_A)_{j,j}}{(\Sigma_A)_{i,i}} \left(\mathbf{u}_A^{(i)}\right)^\top \mathbf{u}_A^{(j)} = \begin{cases} 1 & , \text{if } i = j \\ 0 & , \text{if } i \neq j \end{cases},$$

where the second equality follows from $B^\top B \mathbf{u}_A^{(j)} = AA^\top \mathbf{u}_A^{(j)} = U_A \Sigma_A \Sigma_A^\top U_A^\top \mathbf{u}_A^{(j)} = (\Sigma_A)_{j,j}^2 \mathbf{u}_A^{(j)}$. If $r < d$, we let $\mathbf{u}_B^{(r+1)}, \mathbf{u}_B^{(r+2)}, \ldots, \mathbf{u}_B^{(d)}$ be an arbitrary completion of $\mathbf{u}_B^{(1)}, \mathbf{u}_B^{(2)}, \ldots, \mathbf{u}_B^{(r)}$ to an orthonormal basis for $\mathbb{R}^d$. Define $U_B \in \mathbb{R}^{d,d}$ such that its i'th column vector is $\mathbf{u}_B^{(i)}$, and let $\Sigma_B \in \mathbb{R}_{\geq 0}^{d,k}$ be rectangular-diagonal with $(\Sigma_B)_{i,i} = (\Sigma_A)_{i,i}$ for $i \in [r]$, and $(\Sigma_B)_{i,i} = 0$ for $i > r$. Since both U_B and U_A are orthogonal matrices, the proof concludes by showing that $B = U_B \Sigma_B U_A^\top$. By the definitions of U_B and Σ_B, we may write:

$$U_B \Sigma_B U_A^\top = \sum\nolimits_{i=1}^r (\Sigma_A)_{i,i} \mathbf{u}_B^{(i)} \left(\mathbf{u}_A^{(i)}\right)^\top = B \sum\nolimits_{i=1}^r \mathbf{u}_A^{(i)} \left(\mathbf{u}_A^{(i)}\right)^\top .$$

Notice that $U_A \Sigma_A \Sigma_A^\top U_A^\top$ is an eigenvalue decomposition of $AA^\top$, and hence of $B^\top B$. Therefore, $\ker(B) = \ker(B^\top B) = \mathrm{span}\{\mathbf{u}_A^{(r+1)}, \mathbf{u}_A^{(r+2)}, \ldots, \mathbf{u}_A^{(k)}\}$. With this in hand, for any vector $\mathbf{x} \in \mathbb{R}^k$ we have that:

$$B\mathbf{x} = BU_A U_A^\top \mathbf{x} = B \sum_{i=1}^r \mathbf{u}_A^{(i)} \left(\mathbf{u}_A^{(i)}\right)^\top \mathbf{x} .$$

Therefore, we may conclude $B = B \sum_{i=1}^r \mathbf{u}_A^{(i)} \left(\mathbf{u}_A^{(i)}\right)^\top = U_B \Sigma_B U_A^\top$. □

Lemma 14. *Let $g : [0, \infty) \to \mathbb{R}$ be a continuously differentiable function, and fix some $t > 0$. If $g(t) < g(0)$, then for any $a \in (g(t), g(0)]$ there exists $t_a \in [0, t)$ such that $g(t_a) = a$ and $\dot{g}(t_a) \leq 0$. Similarly, if $g(t) > g(0)$, then for any $a \in [g(0), g(t))$ there exists $t_a \in [0, t)$ such that $g(t_a) = a$ and $\dot{g}(t_a) \geq 0$.*

Proof. Let $t > 0$ be such that $g(t) < g(0)$, and fix some $a \in (g(t), g(0)]$. Define $t_a := \max\{t' : t' \leq t \text{ and } g(t') = a\}$. Continuity of $g(\cdot)$, along with the intermediate value theorem, imply that t_a is well defined (maximum of a closed non-empty set bounded from above). Assume by contradiction that $\dot{g}(t_a) > 0$. Then, $g(\cdot)$ is monotonically increasing on some neighborhood of t_a. Thus, by the intermediate value theorem, there exists $t' \in (t_a, t)$ such that $g(t') = a$, in contradiction to the definition of t_a. An identical argument establishes the analogous result for the case $g(t) > g(0)$. □

Lemma 15. *Let $g : [0, \infty) \to \mathbb{R}$ be a non-negative differentiable function. Assume there exist constants $a, b > 0$ such that $\int_{t'=0}^t g(t')dt' \leq a$ and $\dot{g}(t) \leq b$ for all $t \geq 0$. Then, $lim_{t \to \infty} g(t) = 0$.*

Proof. By way of contradiction let us assume that $g(t)$ does not converge to 0. Let $\epsilon > 0$ be such that for all $M > 0$ there exists $t > M$ with $g(t) > \epsilon$.

We claim that for all $M, \epsilon' > 0$ there exists $t > M$ such that $g(t) < \epsilon'$. Otherwise, we have a contradiction to the bound on the integral of $g(\cdot)$. Combined with our assumption, this means that for all $M > 0$ we can find an interval $[t_1, t_2]$, with $t_1 > M$, where $g(t)$ transitions from $\frac{\epsilon}{2}$ to ϵ. We now examine one such interval. Formally, for t_0 with $g(t_0) < \frac{\epsilon}{2}$, we define:

$$t_2 := \min \{t | t \geq t_0 \text{ and } g(t) = \epsilon\} \quad , \quad t_1 := \max \{t | t \leq t_2 \text{ and } g(t) = \epsilon/2\} \; .$$

Due to the fact that $g(\cdot)$ is continuous, t_2 and t_1 are well defined as they are the minimum and maximum, respectively, of closed non-empty sets bounded from below and above, respectively. Furthermore, notice that $t_0 < t_1 < t_2$. From the mean value theorem and the bound on the derivative of $g(\cdot)$ we have $t_2 - t_1 \geq \epsilon/2b$. Since $g(t) \geq \epsilon/2$ over the interval $[t_1, t_2]$, this gives us $\int_{t'=t_1}^{t_2} g(t')dt' \geq \epsilon^2/4b$. Recall there are infinitely many such occurrences, implying that $\int_{t'=0}^{\infty} g(t')dt' = \infty$, in contradiction to the bound on the integral. □

Lemma 16. *Let $t_0 > 0$ and $g : [t_0, \infty) \to \mathbb{R}$ be a continuously differentiable function such that there exists $T > t_0$ for which $g(t)$ is positive over $[t_0, T)$. Assume that $|\dot{g}(t)| \leq a + b \cdot g(t)^\gamma$ for all $t \in [t_0, T]$, with $a, b > 0$, and $\gamma \geq 1$. Then:*

- *If $\gamma = 1$:*
$$g(T) \geq \frac{a + b \cdot g(t_0)}{b} \cdot e^{-b(T - t_0)} - \frac{a}{b} \, .$$
- *Otherwise, if $\gamma > 1$:*
$$g(T) \geq \frac{1}{b^{1/\gamma} \left[b^{1/\gamma}(\gamma - 1)(T - t_0) + \left(a^{1/\gamma} + b^{1/\gamma} \cdot g(t_0) \right)^{1-\gamma} \right]^{1/(\gamma-1)}} - \left(\frac{a}{b} \right)^{1/\gamma} \, .$$

Proof. Since $g(\cdot)$ is continuous over $[t_0, \infty)$, and positive over $[t_0, T)$, it is non-negative at T. In the case where $\gamma = 1$, we have that $\dot{g}(t) \geq -a - b \cdot g(t)$. Dividing both sides by $a + b \cdot g(t)$ (positive since $a, b > 0$ and $g(t) \geq 0$), and integrating over $t \in [t_0, T]$, we have that:

$$\frac{1}{b} \left[\ln (a + b \cdot g(T)) - \ln (a + b \cdot g(t_0)) \right] = \int_{t=t_0}^{T} \frac{\dot{g}(t)}{a + b \cdot g(t)} dt \geq -(T - t_0) \, .$$

The lower bound on $g(T)$ readily follows by rearranging the inequality above.

Suppose $\gamma > 1$. We begin by showing that $a + b \cdot g(t)^\gamma \leq (a^{1/\gamma} + b^{1/\gamma} \cdot g(t))^\gamma$ whenever $g(t) \geq 0$. To see it is so, notice that for any $\mathbf{x} := (x_1, \ldots, x_d) \in \mathbb{R}^d_{\geq 0}$ it holds that $\sum_{i=1}^d x_i^\gamma \leq (\sum_{i=1}^d x_i)^\gamma$. This is directly implied from the following norm inequality: $\|\mathbf{x}\|_\gamma := (\sum_{i=1}^d x_i^\gamma)^{1/\gamma} \leq \|\mathbf{x}\|_1 := \sum_{i=1}^d x_i$. Thus, for $t \in [t_0, T]$ it holds that $|\dot{g}(t)| \leq (a^{1/\gamma} + b^{1/\gamma} \cdot g(t))^\gamma$, and in particular:

$$\dot{g}(t) \geq - \left(a^{1/\gamma} + b^{1/\gamma} \cdot g(t) \right)^\gamma \, .$$

Dividing by $(a^{1/\gamma} + b^{1/\gamma} \cdot g(t))^\gamma$ (positive since $a, b > 0$ and $g(t) \geq 0$), and integrating over $t \in [t_0, T]$:

$$\int_{t=t_0}^{T} \frac{\dot{g}(t)}{\left(a^{1/\gamma} + b^{1/\gamma} \cdot g(t) \right)^\gamma} dt \geq -(T - t_0)$$
$$\Longrightarrow \frac{1}{b^{1/\gamma}(1 - \gamma)} \left[\left(a^{1/\gamma} + b^{1/\gamma} \cdot g(T) \right)^{1-\gamma} - \left(a^{1/\gamma} + b^{1/\gamma} \cdot g(t_0) \right)^{1-\gamma} \right] \geq -(T - t_0) \, .$$

Rearranging the inequality above establishes the desired result. □

Lemma 17. *Let $\mathcal{D}$ be an open domain in Euclidean space, and let $f : \mathcal{D} \to \mathbb{R}$ be a continuously differentiable function. Let $\theta : [0, \infty) \to \mathcal{D}$ be a curve born from gradient flow over $f(\cdot)$:*

$$\theta(0) = \theta_0 \in \mathcal{D} \quad , \quad \dot{\theta}(t) := \tfrac{d}{dt}\theta(t) = -\nabla f(\theta(t)) \; , \; t \geq 0 \, .$$

Then, $f(\theta(t))$ is monotonically non-increasing with respect to t.

Proof. The proof immediately follows from the fact that for any $t \in [0, \infty)$:

$$\frac{d}{dt} f(\theta(t)) = \left\langle \nabla f(\theta(t)), \dot{\theta(t)} \right\rangle = \langle \nabla f(\theta(t)), -\nabla f(\theta(t)) \rangle = - \|\nabla f(\theta(t))\|_2^2 \leq 0 \, .$$

□

G.3 Proof of Proposition 1

For a quasi-norm $\|\cdot\|$, the weakened triangle inequality (see Footnote 2) implies that there exists a constant $c_{\|\cdot\|} \geq 1$ for which

$$\begin{aligned} \|W\| &\geq \frac{1}{c_{\|\cdot\|}} \left\|(W)_{1,1}\mathbf{e}_1\mathbf{e}_1^\top\right\| - \left\|W - (W)_{1,1}\mathbf{e}_1\mathbf{e}_1^\top\right\| \\ &= |(W)_{1,1}| \frac{\left\|\mathbf{e}_1\mathbf{e}_1^\top\right\|}{c_{\|\cdot\|}} - \left\|\mathbf{e}_2\mathbf{e}_1^\top + \mathbf{e}_1\mathbf{e}_3^\top\right\| , \end{aligned} \tag{28}$$

for any $W \in \mathcal{S}$. Fix some $\epsilon > 0$ and define $M_{\|\cdot\|,\epsilon} := \{(W)_{1,1} \in \mathbb{R} : \|W\| \leq \inf_{W' \in \mathcal{S}} \|W'\| + \epsilon,\ \ W \in \mathcal{S}\}$, the set of $(W)_{1,1}$ values corresponding to ϵ-minimizers of $\|\cdot\|$. The first part of the proposition thus boils down to showing $M_{\|\cdot\|,\epsilon}$ is bounded. By Equation (28), there exist a $C > 0$ such that $|(W)_{1,1}| > C$ means $\|W\| > \inf_{W' \in \mathcal{S}} \|W'\| + \epsilon$. Hence, $M_{\|\cdot\|,\epsilon} \subset I_{\|\cdot\|,\epsilon} := [-C, C]$.

If in addition $\|\cdot\|$ is a Schatten-p quasi-norm for $p \in (0, \infty]$, we now show that W is its minimizer over $\mathcal{S}$ if and only if $(W)_{1,1} = 0$. Let $W_x \in \mathcal{S}$ denote the solution matrix with $(W_x)_{1,1} = x$ for $x \in \mathbb{R}$. The singular values of an arbitrary such W_x are:

$$\{\sigma_1(W_x), \sigma_2(W_x)\} = \left\{ \left| \left(x + \sqrt{x^2+4}\right)/2 \right|, \left| \left(x - \sqrt{x^2+4}\right)/2 \right| \right\} . \tag{29}$$

Starting with $p = \infty$, the corresponding norm is the spectral norm $\|W_x\|_{S_\infty} := \sigma_1(W_x)$. When $x = 0$, we have that $\sigma_1(W_0) = 1$. If $x > 0$, then $\sigma_1(W_x) = (x + \sqrt{x^2+4})\ /\ 2 > 1$. Similarly, if $x < 0$, then $\sigma_1(W_x) = (-x + \sqrt{x^2+4})\ /\ 2 > 1$. Therefore, $\|W_x\|_{S_\infty}$ attains its minimal value of 1 if and only if $x = 0$.

Moving to the case of $p \in (0, \infty)$, the corresponding quasi-norm is $\|W_x\|_{S_p} := (\sigma_1(W_x)^p + \sigma_2(W_x)^p)^{\frac{1}{p}}$. We now examine $\|W_x\|_{S_p}^p$ for $x > 0$:

$$\|W_x\|_{S_p}^p = \left(\frac{x + \sqrt{x^2+4}}{2} \right)^p + \left(\frac{-x + \sqrt{x^2+4}}{2} \right)^p .$$

Differentiating with respect to x, we arrive at:

$$\begin{aligned} &\frac{p}{2^p} \left(\left(x + \sqrt{x^2+4}\right)^{p-1} \left(1 + \frac{x}{\sqrt{x^2+4}}\right) + \left(-x + \sqrt{x^2+4}\right)^{p-1} \left(-1 + \frac{x}{\sqrt{x^2+4}}\right) \right) \\ &> \frac{p}{2^p} \left(\left(x + \sqrt{x^2+4}\right)^{p-1} - \left(-x + \sqrt{x^2+4}\right)^{p-1} \right) \\ &> 0 , \end{aligned}$$

where in the first transition we used the fact that both $\left(x + \sqrt{x^2+4}\right)^{p-1}$ and $\left(-x + \sqrt{x^2+4}\right)^{p-1}$ are positive (as well as x). It then directly follows that $\|W_x\|_{S_p}^p$ and thus $\|W_x\|_{S_p}$ are monotonically increasing with respect to x on $(0, \infty)$.

Similar arguments show that when $x < 0$ the Schatten-p quasi-norm of W_x is monotonically decreasing with respect to x, implying that $\|W_x\|_{S_p}$ is minimized if and only if $x = 0$.[20] □

G.4 Proof of Proposition 2

As in the proof of Proposition 1 (Subappendix G.3), we denote by $W_x \in \mathcal{S}$ the solution matrix with $(W_x)_{1,1} = x$. We begin by analyzing the behavior of $\sigma_1(W_x)$ and $\sigma_2(W_x)$ with respect to x. When $x = 0$ the singular values are simply $\sigma_1(W_0) = \sigma_2(W_0) = 1$. When x is positive, the singular values may be written as:

$$\sigma_1(W_x) = \frac{x + \sqrt{x^2+4}}{2} \quad , \quad \sigma_2(W_x) = \frac{-x + \sqrt{x^2+4}}{2} .$$

[20] The claim relies on the fact that the Schatten-p quasi-norm of W_x is continuous with respect to x for all $p \in (0, \infty)$. We note, however, that quasi-norms in general may be discontinuous.

Taking the derivative with respect to x, we arrive at:

$$\frac{d}{dx}\sigma_1(W_x) = \frac{1}{2} + \frac{x}{2\sqrt{x^2+4}} \quad , \quad \frac{d}{dx}\sigma_2(W_x) = -\frac{1}{2} + \frac{x}{2\sqrt{x^2+4}} \, .$$

Since $x > 0$, we have that $d/dx \ \sigma_1(W_x) > 0$ and $d/dx \ \sigma_2(W_x) < 0$. In other words, $\sigma_1(W_x)$ is monotonically increasing, while $\sigma_2(W_x)$ is monotonically decreasing, when $x > 0$. It can easily be verified that $\sigma_1(W_x)$ and $\sigma_2(W_x)$ are even functions of x, *i.e.* $\sigma_1(W_x) = \sigma_1(W_{-x})$ and $\sigma_2(W_x) = \sigma_2(W_{-x})$. It then follows that $\sigma_1(W_x)$ is monotonically decreasing (conversely $\sigma_2(W_x)$ is monotonically increasing) when $x < 0$. Noticing that $\lim_{x\to\infty}\sigma_1(W_x) = \infty$ and $\lim_{x\to\infty}\sigma_2(W_x) = 0$ (accordingly $\lim_{x\to-\infty}\sigma_1(W_x) = \infty$ and $\lim_{x\to-\infty}\sigma_2(W_x) = 0$), we have a characterization of the behavior of $\sigma_1(W_x)$ and $\sigma_2(W_x)$.

We are now in a position to obtain the desired results for effective and infimal ranks. The effective rank (Definition 1) of W_x can be written as

$$\mathrm{erank}(W_x) = \exp\left\{ H\left(\frac{\sigma_1(W_x)}{\sigma_1(W_x)+\sigma_2(W_x)}, \frac{\sigma_2(W_x)}{\sigma_1(W_x)+\sigma_2(W_x)} \right) \right\} \, .$$

The binary entropy function is bounded by $\ln(2)$, hence, the effective rank over $\mathcal{S}$ is bounded by 2. This upper bound is attained at $x = 0$. According to the singular values analysis, when $|x| \to \infty$ we have that $\rho_1(W_x)$ monotonically increases towards 1, starting from the value $\rho_1(W_0) = \frac{1}{2}$. Noticing that this implies the entropy function and effective rank monotonically decrease towards 0 and 1, respectively, completes the effective rank analysis.

Next, we analyze the infimal rank of $\mathcal{S}$ and the distance of W_x from that infimal rank. The distance of W_x from $\mathcal{M}_1$ is $D(W_x, \mathcal{M}_1) = \sigma_2(W_x)$. Since $\lim_{x\to\infty}\sigma_2(W_x) = 0$, we have $D(\mathcal{S}, \mathcal{M}_1) = 0$. Clearly $D(\mathcal{S}, \mathcal{M}_0) > 0$, leading to the conclusion that the infimal rank of $\mathcal{S}$ is 1. Finally, the analysis of $\sigma_2(W_x)$ directly implies that the distance of W_x from the infimal rank of $\mathcal{S}$ is maximized when $x = 0$, monotonically tending to 0 as $|x| \to \infty$. □

G.5 Proof of Theorem 1

In the following, as stated in Subappendix G.1, for results that hold for all $t \geq 0$ or when clear from the context, we omit the time index t. Furthermore, we denote the entries of the product matrix $W_{L:1}$ by $\{w_{i,j}\}_{i,j\in[2]}$.

We begin by deriving loss-dependent bounds for $|w_{1,1}|$, $\sigma_1(W_{L:1})$ and $\sigma_2(W_{L:1})$. Writing the loss explicitly:

$$\ell(W_{L:1}) = \frac{1}{2}\left[(w_{1,2}-1)^2 + (w_{2,1}-1)^2 + w_{2,2}^2 \right] \, ,$$

we can upper bound each of the non-negative terms separately. Multiplying by 2 and taking the square root of both sides yields:

$$|w_{2,2}| \leq \sqrt{2\ell(W_{L:1})} \quad , \quad |w_{1,2}-1| \leq \sqrt{2\ell(W_{L:1})} \quad , \quad |w_{2,1}-1| \leq \sqrt{2\ell(W_{L:1})} \, . \tag{30}$$

The following lemma characterizes the relation between $|w_{1,1}|$ and the loss.

Lemma 18. *Suppose* $\ell(W_{L:1}) < \frac{1}{2}$. *Then:*

$$|w_{1,1}| > \frac{(1-\sqrt{2\ell(W_{L:1})})^2}{\sqrt{2\ell(W_{L:1})}} = \frac{1}{\sqrt{2\ell(W_{L:1})}} - 2 + \sqrt{2\ell(W_{L:1})} \, .$$

Proof. From Lemma 6, the determinant of $W_{L:1}$ does not change signs and remains positive, *i.e.*:

$$\det(W_{L:1}) = w_{1,1}w_{2,2} - w_{1,2}w_{2,1} > 0 \, . \tag{31}$$

Under the assumption that $\ell(W_{L:1}) < \frac{1}{2}$, both $w_{1,2}$ and $w_{2,1}$ are positive and lie inside the open interval $(0, 2)$. Since the determinant is positive, $w_{2,2} \neq 0$ and $w_{1,1}w_{2,2} > 0$ must hold. Rearranging Equation (31), we may therefore write $|w_{1,1}w_{2,2}| > w_{1,2}w_{2,1}$. Dividing both sides by $|w_{2,2}|$ and applying the bounds from Equation (30) completes the proof:

$$|w_{1,1}| > \frac{(1-\sqrt{2\ell(W_{L:1})})^2}{\sqrt{2\ell(W_{L:1})}} = \frac{1}{\sqrt{2\ell(W_{L:1})}} - 2 + \sqrt{2\ell(W_{L:1})} \, .$$

□

An immediate consequence of the lemma above is that decreasing the loss towards zero drives $|w_{1,1}|$ towards infinity.

With this bound in hand, Lemma 19 below establishes bounds on the singular values of $W_{L:1}$. In turn, they will allow us to obtain the necessary results for effective rank (Definition 1) and distance from infimal rank of $\mathcal{S}$ (Definition 2).

Lemma 19. *The singular values of $W_{L:1}$ fulfill:*

$$\sigma_1(W_{L:1}) \geq |w_{1,1}| - \sqrt{2\ell(W_{L:1})} \quad , \quad \sigma_2(W_{L:1}) \leq 3\sqrt{2\ell(W_{L:1})} \,. \tag{32}$$

Furthermore, if $\ell(W_{L:1}) < \frac{1}{2}$, then:

$$\sigma_1(W_{L:1}) \geq \frac{1}{\sqrt{2\ell(W_{L:1})}} - 2 \,. \tag{33}$$

Proof. Define $W_{\mathcal{S}} := \begin{pmatrix} w_{1,1} & 1 \\ 1 & 0 \end{pmatrix}$, the orthogonal projection of $W_{L:1}$ onto the solution set $\mathcal{S}$. By Corollary 8.6.2 in [33] we have that:

$$|\sigma_i(W_{L:1}) - \sigma_i(W_{\mathcal{S}})| \leq \|W_{L:1} - W_{\mathcal{S}}\|_F = \sqrt{2\ell(W_{L:1})} \quad , \; i = 1, 2 \,. \tag{34}$$

One can easily verify that $W_{\mathcal{S}}$ is a symmetric indefinite matrix with eigenvalues

$$\{\lambda_1(W_{\mathcal{S}}), \lambda_2(W_{\mathcal{S}})\} = \left\{ \left(w_{1,1} + \sqrt{w_{1,1}^2 + 4}\right) / \, 2 \,, \; \left(w_{1,1} - \sqrt{w_{1,1}^2 + 4}\right) / \, 2 \right\} \,.$$

Suppose that $w_{1,1} \geq 0$. We thus have:

$$\sigma_1(W_{\mathcal{S}}) = \max_{i=1,2} |\lambda_i(W_{\mathcal{S}})| = \frac{w_{1,1} + \sqrt{w_{1,1}^2 + 4}}{2} \geq |w_{1,1}| \,,$$

and

$$\begin{aligned} \sigma_2(W_{\mathcal{S}}) &= \min_{i=1,2} |\lambda_i(W_{\mathcal{S}})| \\ &= \frac{\sqrt{w_{1,1}^2 + 4} - w_{1,1}}{2} \\ &= \frac{2}{\sqrt{w_{1,1}^2 + 4} + w_{1,1}} \\ &\leq \frac{2}{2 + w_{1,1}} \,, \end{aligned}$$

where in the third transition we made use of the identity $a - b = \frac{a^2 - b^2}{a + b}$ for $a, b \in \mathbb{R}$ such that $a + b \neq 0$. If $\ell(W_{L:1}) \geq \frac{1}{2}$, it holds that $\sigma_2(W_{\mathcal{S}}) \leq 2/(2 + w_{1,1}) \leq 1 \leq 2\sqrt{2\ell(W_{L:1})}$. Otherwise, we may apply the lower bound on $w_{1,1}$ (Lemma 18) and conclude that $\sigma_2(W_{\mathcal{S}}) \leq 2\sqrt{2\ell(W_{L:1})}$ for any loss value. Having established that $\sigma_1(W_{\mathcal{S}}) \geq |w_{1,1}|$ and $\sigma_2(W_{\mathcal{S}}) \leq 2\sqrt{2\ell(W_{L:1})}$, Equation (34) completes the proof of Equation (32). It remains to see that if $\ell(W_{L:1}) < \frac{1}{2}$, from the lower bound on $w_{1,1}$ (Lemma 18), Equation (33) immediately follows.

By similar arguments, Equations (32) and (33) hold for $w_{1,1} < 0$ as well. □

G.5.1 Proof of Equation (8) (lower bound for quasi-norm)

We turn to lower bound the quasi-norm of the product matrix. It holds that:

$$\|W_{L:1}\| \geq \frac{1}{c_{\|\cdot\|}} \left\|w_{1,1}\mathbf{e}_1\mathbf{e}_1^\top\right\| - \left\|W_{L:1} - w_{1,1}\mathbf{e}_1\mathbf{e}_1^\top\right\| \,, \tag{35}$$

where $c_{\|\cdot\|} \geq 1$ is a constant for which $\|\cdot\|$ satisfies the weakened triangle inequality (see Footnote 2). We now assume that $\ell(W_{L:1}) < \frac{1}{2}$. Later this assumption will be lifted, providing a bound that holds for all loss values. Subsequent applications of the weakened triangle inequality, together with homogeneity of $\|\cdot\|$ and the bounds on the entries of $W_{L:1}$ (Equation (30)), give:

$$\begin{aligned}\left\|W_{L:1} - w_{1,1}\mathbf{e}_1\mathbf{e}_1^\top\right\| &\leq c_{\|\cdot\|}|w_{2,2}| \left\|\mathbf{e}_2\mathbf{e}_2^\top\right\| + c_{\|\cdot\|}^2 \left(|w_{2,1}| \left\|\mathbf{e}_2\mathbf{e}_1^\top\right\| + |w_{1,2}| \left\|\mathbf{e}_1\mathbf{e}_2^\top\right\|\right) \\ &\leq c_{\|\cdot\|}\sqrt{2\ell(W_{L:1})} \left\|\mathbf{e}_2\mathbf{e}_2^\top\right\| + c_{\|\cdot\|}^2 \left(1 + \sqrt{2\ell(W_{L:1})}\right) \left(\left\|\mathbf{e}_2\mathbf{e}_1^\top\right\| + \left\|\mathbf{e}_1\mathbf{e}_2^\top\right\|\right) \\ &\leq c_{\|\cdot\|} \left\|\mathbf{e}_2\mathbf{e}_2^\top\right\| + 2c_{\|\cdot\|}^2 \left(\left\|\mathbf{e}_2\mathbf{e}_1^\top\right\| + \left\|\mathbf{e}_1\mathbf{e}_2^\top\right\|\right) \\ &\leq 2c_{\|\cdot\|}^2 \left(\left\|\mathbf{e}_2\mathbf{e}_2^\top\right\| + \left\|\mathbf{e}_2\mathbf{e}_1^\top\right\| + \left\|\mathbf{e}_1\mathbf{e}_2^\top\right\|\right) \,.\end{aligned}$$

Plugging the inequality above and the lower bound on $|w_{1,1}|$ (Lemma 18) into Equation (35), we have:

$$\begin{aligned}\|W_{L:1}\| &\geq \frac{\left\|\mathbf{e}_1\mathbf{e}_1^\top\right\|}{c_{\|\cdot\|}} \left(\frac{1}{\sqrt{2\ell(W_{L:1})}} - 2 + \sqrt{2\ell(W_{L:1})}\right) - 2c_{\|\cdot\|}^2 \left(\left\|\mathbf{e}_2\mathbf{e}_2^\top\right\| + \left\|\mathbf{e}_2\mathbf{e}_1^\top\right\| + \left\|\mathbf{e}_1\mathbf{e}_2^\top\right\|\right) \\ &\geq \frac{\left\|\mathbf{e}_1\mathbf{e}_1^\top\right\|}{c_{\|\cdot\|}} \frac{1}{\sqrt{2\ell(W_{L:1})}} - 2\frac{\left\|\mathbf{e}_1\mathbf{e}_1^\top\right\|}{c_{\|\cdot\|}} - 2c_{\|\cdot\|}^2 \left(\left\|\mathbf{e}_2\mathbf{e}_2^\top\right\| + \left\|\mathbf{e}_2\mathbf{e}_1^\top\right\| + \left\|\mathbf{e}_1\mathbf{e}_2^\top\right\|\right) \\ &\geq \frac{\left\|\mathbf{e}_1\mathbf{e}_1^\top\right\|}{c_{\|\cdot\|}} \frac{1}{\sqrt{2\ell(W_{L:1})}} - 2c_{\|\cdot\|}^2 \left(\left\|\mathbf{e}_1\mathbf{e}_1^\top\right\| + \left\|\mathbf{e}_2\mathbf{e}_2^\top\right\| + \left\|\mathbf{e}_2\mathbf{e}_1^\top\right\| + \left\|\mathbf{e}_1\mathbf{e}_2^\top\right\|\right) \,.\end{aligned}$$

Since $\|W_{L:1}\|$ is trivially lower bounded by zero, defining the constants

$$a_{\|\cdot\|} := \frac{\left\|\mathbf{e}_1\mathbf{e}_1^\top\right\|}{\sqrt{2}c_{\|\cdot\|}} \;,\; b_{\|\cdot\|} := \max\left\{\sqrt{2}a_{\|\cdot\|}, 8c_{\|\cdot\|}^2 \max_{i,j\in\{1,2\}} \left\|\mathbf{e}_i\mathbf{e}_j^\top\right\|\right\} ,$$

allows us, on the one hand, to arrive at a bound of the form:

$$\|W_{L:1}\| \geq a_{\|\cdot\|} \cdot \frac{1}{\sqrt{\ell(W_{L:1})}} - b_{\|\cdot\|} \,,$$

and on the other hand, to lift our previous assumption on the loss: when $\ell(W_{L:1}) \geq \frac{1}{2}$ the bound is vacuous, *i.e.* non-positive and trivially holds. Noticing this is exactly Equation (8) (recall we omitted the time index t), concludes the first part of the proof.

G.5.2 Proof of Equation (9) (upper bound for effective rank)

During the following effective rank (Definition 1) analysis we operate under the assumption of $\ell(W_{L:1}) < \frac{1}{32}$. We later remove this assumption, delivering a bound that holds for all loss values. Making use of the obtained bounds on $\sigma_1(W_{L:1})$ and $\sigma_2(W_{L:1})$ (Lemma 19) we arrive at:

$$\begin{aligned}\rho_1(W_{L:1}) &= \frac{\sigma_1(W_{L:1})}{\sigma_1(W_{L:1}) + \sigma_2(W_{L:1})} \\ &\geq \frac{\sigma_1(W_{L:1})}{\sigma_1(W_{L:1}) + 3\sqrt{2\ell(W_{L:1})}} \\ &= 1 - \frac{3\sqrt{2\ell(W_{L:1})}}{\sigma_1(W_{L:1}) + 3\sqrt{2\ell(W_{L:1})}} \\ &\geq 1 - \frac{3\sqrt{2\ell(W_{L:1})}}{\frac{1}{\sqrt{2\ell(W_{L:1})}} - 2 + 3\sqrt{2\ell(W_{L:1})}} \\ &= 1 - \frac{6\ell(W_{L:1})}{6\ell(W_{L:1}) - 2\sqrt{2\ell(W_{L:1})} + 1} \,.\end{aligned}$$

Given our assumption on the loss, we have $1 - 2\sqrt{2\ell((W_{L:1}))} \geq \frac{1}{2}$ and thus

$$\rho_2(W_{L:1}) = 1 - \rho_1(W_{L:1}) \leq \frac{6\ell(W_{L:1})}{6\ell(W_{L:1}) + \frac{1}{2}} \leq 12\ell(W_{L:1}) \,. \tag{36}$$

Let $h\left(\rho_2(W_{L:1})\right) := -\rho_2(W_{L:1}) \cdot \ln\left(\rho_2(W_{L:1})\right) - \left(1 - \rho_2(W_{L:1})\right) \cdot \ln\left(1 - \rho_2(W_{L:1})\right)$ denote the binary entropy function, and recall that the effective rank of $W_{L:1}$ is defined to be $\mathrm{erank}(W_{L:1}) :=$

$\exp\{h\left(\rho_2(W_{L:1})\right)\}$. The exponent function is convex and therefore upper bounded on the interval $[0, \ln(2)]$ by the linear function that intersects it at these points. Formally, for $x \in [0, \ln(2)]$ it holds that $\exp(x) \leq 1 + \frac{1}{\ln(2)} x$, yielding the following bound:

$$\mathrm{erank}(W_{L:1}) \leq 1 + \frac{1}{\ln(2)} \cdot h\left(\rho_2(W_{L:1})\right) .$$

By Lemma 9 we have that $h\left(\rho_2(W_{L:1})\right) \leq 2\sqrt{\rho_2(W_{L:1})}$. Combined with Equation (36), since $\inf_{W' \in \mathcal{S}} \mathrm{erank}(W') = 1$ (Proposition 2), this leads to:

$$\mathrm{erank}(W_{L:1}) \leq \inf_{W' \in \mathcal{S}} \mathrm{erank}(W') + \frac{2\sqrt{12}}{\ln(2)} \cdot \sqrt{\ell(W_{L:1})} .$$

Recall that the time index t is omitted, and the result holds for all $t \geq 0$, *i.e.* this is exactly Equation (9). To remove our assumption on the loss, notice that when $\ell(W_{L:1}) \geq \frac{1}{32}$ the bound is trivial, as the right-hand side is greater than 2, which is the maximal effective rank (for a 2 × 2 matrix).

G.5.3 Proof of Equation (10) (upper bound for distance from infimal rank)

According to Proposition 2, the infimal rank of $\mathcal{S}$ is 1. The quantity we seek to upper bound is therefore $D(W_{L:1}(t), \mathcal{M}_1) = \sigma_2(W_{L:1}(t))$. By Equation (32) in Lemma 19, for all $t \geq 0$ we have

$$D(W_{L:1}(t), \mathcal{M}_1) \leq 3\sqrt{2} \cdot \sqrt{\ell(t)} ,$$

completing the proof. □

G.6 Proof of Proposition 3

Define W_{-1} to be the matrix obtained from W by multiplying its first row by -1. On the one hand, symmetry around the origin implies that W_{-1} and W follow the same distribution. On the other hand, $\det(W_{-1}) = -\det(W)$. Due to the fact that the set of matrices with zero determinant has probability 0 under continuous distributions (see, *e.g.*, Remark 2.5 in [37]), we may conclude $\Pr(\det(W) > 0) = \Pr(\det(W) < 0) = 0.5$.

For $W_1, W_2, \ldots, W_L$ random matrices drawn independently, let $l \in [L]$ be the index such that $\Pr(\det(W_l) > 0) = 0.5$. Since $Pr(\det(W_{l'}) = 0) = 0$ for any $l' \in [L]$, the proof readily follows from determinant multiplicativity and the law of total probability:

$$\begin{aligned}
\Pr\left(\det(W_L W_{L-1} \cdots W_1) > 0\right) &= \Pr(\det(W_l) > 0) \cdot \Pr\left(\Pi_{i \neq l} \det(W_i) > 0\right) \\
&\quad + \Pr(\det(W_l) < 0) \cdot \Pr\left(\Pi_{i \neq l} \det(W_i) < 0\right) \\
&= \frac{1}{2}\left[\Pr\left(\Pi_{i \neq l} \det(W_i) > 0\right) + \Pr\left(\Pi_{i \neq l} \det(W_i) < 0\right)\right] \\
&= 0.5 .
\end{aligned}$$

An identical computation yields $\Pr\left(\det(W_L W_{L-1} \cdots W_1) < 0\right) = 0.5$. □

G.7 Proof of Proposition 4

The proof makes use of the following lemma, proven in Subappendix G.7.1.

Lemma 20. *Consider the setting of Theorem 1 (arbitrary depth $L \in \mathbb{N}$) in the special case of an initial product matrix $W_{L:1}(0) = \alpha \cdot I$, where I stands for identity matrix and $\alpha \in (0, 1]$. Then, $W_{L:1}(t)$ is positive definite for all $t \geq 0$.*

With Lemma 20 in place, we may derive the exact differential equations governing the product matrix in our setting of depth $L = 2$. Then, a detailed analysis of the dynamics will yield convergence of the loss to global minimum, *i.e.* $\lim_{t \to \infty} \ell(t) = 0$. As usual, we omit the time index t when stating results for all t or when clear from the context.

According to Lemma 20, the product matrix $W_{L:1}$ is symmetric and positive definite. Thus, we may write the loss and its gradient with respect to $W_{L:1}$ as:

$$\ell(W_{L:1}) = \frac{1}{2}\left[w_{2,2}^2 + 2(w_{1,2} - 1)^2\right] \quad , \quad \nabla \ell(W_{L:1}) = \begin{pmatrix} 0 & w_{1,2} - 1 \\ w_{1,2} - 1 & w_{2,2} \end{pmatrix} , \qquad (37)$$

where $\{w_{i,j}\}_{i,j \in [2]}$ are the entries of $W_{L:1}$. Since the factors W_1 and W_2 are balanced at initialization (Equation (5)), the differential equation governing the product matrix (Lemma 4) for depth $L = 2$

gives:

$$\begin{aligned}\dot{W}_{L:1} &= -\left[W_{L:1}W_{L:1}^\top\right]^{\frac{1}{2}} \cdot \nabla\ell(W_{L:1}) - \nabla\ell(W_{L:1}) \cdot \left[W_{L:1}^\top W_{L:1}\right]^{\frac{1}{2}} \\ &= -W_{L:1}\nabla\ell(W_{L:1}) - \nabla\ell(W_{L:1})W_{L:1}\,,\end{aligned} \tag{38}$$

where the transition is by positive definiteness of $W_{L:1}$. Writing the differential equation of each entry separately, we have:

$$\begin{aligned}\dot{w}_{1,1} &= 2w_{1,2}(1-w_{1,2})\,,\\ \dot{w}_{2,2} &= 2w_{1,2}(1-w_{1,2}) - 2w_{2,2}^2\,,\\ \dot{w}_{1,2} &= w_{2,2}(1-2w_{1,2}) + w_{1,1}(1-w_{1,2})\,.\end{aligned} \tag{39}$$

Let us characterize the behavior of these entries throughout time.

Lemma 21. *The following holds for all $t \geq 0$:*

1. $w_{1,1} > 0$ *and is monotonically non-decreasing.*
2. $0 \leq w_{1,2} \leq 1$.
3. $0 < w_{2,2} \leq 1$.

Proof. Since $W_{L:1}$ is positive definite, it follows that $w_{1,1}$ and $w_{2,2}$ are positive. Examining the behavior of $w_{1,2}$ (Equation (39)): on the one hand, when $w_{1,2} = 0$ then $\dot{w}_{1,2} = w_{2,2} + w_{1,1} > 0$, and on the other hand, when $w_{1,2} = 1$ then $\dot{w}_{1,2} = -w_{2,2} < 0$. Because $w_{1,2}$ is initialized at 0, it stays in the interval $[0,1]$. Otherwise, by Lemma 14, we have a contradiction to the positivity of $\dot{w}_{1,2}$ when $w_{1,2} = 0$ or its negativity when $w_{1,2} = 1$. Similarly, if $w_{2,2} > \frac{1}{2}$ we have $\dot{w}_{2,2} < 2w_{1,2}(1-w_{1,2}) - \frac{1}{2} \leq 0$. Since at initialization $w_{2,2}(0) = \alpha \leq 1$, by Lemma 14, it will not go above 1. Lastly, since $w_{1,2}$ is in the interval $[0,1]$, it holds that $\dot{w}_{1,1} \geq 0$, *i.e.* $w_{1,1}$ is monotonically non-decreasing. □

We turn our focus to the derivative of the loss with respect to t:

$$\frac{d}{dt}\ell(W_{L:1}) = \langle \nabla\ell(W_{L:1}), \dot{W}_{L:1}\rangle\,.$$

Plugging in Equation (38) and recalling the fact that $\langle A, B\rangle = \mathrm{Tr}(A^\top B)$ for matrices A, B of the same size:

$$\frac{d}{dt}\ell(W_{L:1}) = -\,\mathrm{Tr}(\nabla\ell(W_{L:1})^\top W_{L:1}\nabla\ell(W_{L:1})) - \mathrm{Tr}(\nabla\ell(W_{L:1})^\top \nabla\ell(W_{L:1})W_{L:1})\,.$$

From the cyclic property of the trace operator and symmetry of $\nabla\ell(W_{L:1})$ (Equation (37)), we arrive at the following expression:

$$\frac{d}{dt}\ell(W_{L:1}) = -2\,\mathrm{Tr}(\nabla\ell(W_{L:1})W_{L:1}\nabla\ell(W_{L:1}))\,.$$

Notice that since $\nabla\ell(W_{L:1})W_{L:1}\nabla\ell(W_{L:1})$ is positive semidefinite the trace is non-negative and $\frac{d}{dt}\ell(W_{L:1}) \leq 0$. That is, the loss is monotonically non-increasing throughout time. Invoking Lemma 11, we can upper bound the derivative by:

$$\frac{d}{dt}\ell(W_{L:1}) \leq -2\lambda_1(W_{L:1}) \cdot \sigma_2(\nabla\ell(W_{L:1}))^2\,, \tag{40}$$

where $\lambda_1(W_{L:1})$ is the maximal eigenvalue of $W_{L:1}$ and $\sigma_2(\nabla\ell(W_{L:1}))$ is the minimal singular value of $\nabla\ell(W_{L:1})$. The maximal eigenvalue of a symmetric matrix is greater than its diagonal entries. Therefore, $\lambda_1(W_{L:1}) \geq w_{1,1}$. Since $w_{1,1}$ is initialized at $\alpha > 0$, and by Lemma 21 is monotonically non-decreasing, we have $\lambda_1(W_{L:1}) \geq \alpha$. Writing the eigenvalues of $\nabla\ell(W_{L:1})$ explicitly:

$$\begin{aligned}\lambda_1(\nabla\ell(W_{L:1})) &= \frac{w_{2,2} + \sqrt{w_{2,2}^2 + 4(1-w_{1,2})^2}}{2}\,,\\ \lambda_2(\nabla\ell(W_{L:1})) &= \frac{w_{2,2} - \sqrt{w_{2,2}^2 + 4(1-w_{1,2})^2}}{2}\,,\end{aligned}$$

we can see that, since $w_{2,2}$ is positive (Lemma 21), $\sigma_2(\nabla\ell(W_{L:1})) = \min_{i=1,2}|\lambda_i(\nabla\ell(W_{L:1}))| = (\sqrt{w_{2,2}^2 + 4(1-w_{1,2})^2} - w_{2,2})/2$. Applying the identity $a - b = \frac{a^2-b^2}{a+b}$ for $a, b \in \mathbb{R}$ such that $a + b \neq 0$, and the bounds on $w_{2,2}$ and $w_{1,2}$ (Lemma 21):

$$\begin{aligned}\sigma_2(\nabla\ell(W_{L:1})) &= \frac{2(1-w_{1,2})^2}{\sqrt{w_{2,2}^2 + 4(1-w_{1,2})^2} + w_{2,2}} \\ &\geq \frac{2(1-w_{1,2})^2}{\sqrt{1 + 4(1-w_{1,2})^2} + 1} \\ &\geq \frac{2(1-w_{1,2})^2}{2\,|(1-w_{1,2})| + 2} \\ &\geq \frac{1}{2}(1-w_{1,2})^2\,,\end{aligned}$$

where in the penultimate transition we bounded the square root of a sum by the sum of square roots. Returning to Equation (40) we have:

$$\frac{d}{dt}\ell(W_{L:1}) \leq -b(1-w_{1,2})^4\,,$$

for $b = \frac{1}{2}\alpha$. We are now in a position to prove that $w_{1,2} \to 1$ as t tends to infinity. Integrating both sides with respect to time:

$$\ell(W_{L:1}(t)) - \ell(W_{L:1}(0)) \leq -b\int_{t'=0}^{t}(1-w_{1,2}(t'))^4 dt'\,.$$

Since $\ell(W_{L:1}(t)) \geq 0$, by rearranging the inequality we may write:

$$\int_{t'=0}^{t}(1-w_{1,2}(t'))^4 dt' \leq \frac{\ell(W_{L:1}(0))}{b}\,.$$

Going back to the differential equation of $\dot{w}_{1,2}$ (Equation (39)), by applying the bounds on $w_{1,2}$ and $w_{2,2}$ (Lemma 21) we have that $\dot{w}_{1,2} \geq -1$. Defining $g(t) := (1-w_{1,2}(t))^4$, it then holds that $\dot{g}(t) \leq 4$. Since $g(\cdot)$ is non-negative and has an upper bounded integral and derivative, from Lemma 15, we can conclude that $lim_{t\to\infty}g(t) = 0$ and $lim_{t\to\infty}w_{1,2}(t) = 1$.

Because $\ell(W_{L:1}(t))$ is monotonically non-increasing, we need only show that for each $\epsilon > 0$ there exists a $t_\epsilon > 0$ such that $\ell(W_{L:1}(t_\epsilon)) < \epsilon$. Having already established that $w_{1,2}(t)$ converges to 1, this amounts to finding a large enough t_ϵ for which $w_{2,2}(t_\epsilon)$ is sufficiently close to 0. Fix some $\epsilon > 0$ and let $\hat{t} > 0$ be such that for all $t \geq \hat{t}$ the following holds:

$$2(1-w_{1,2}(t))^2 < \epsilon \quad , \quad 2w_{1,2}(t)(1-w_{1,2}(t)) < \epsilon\,. \tag{41}$$

Such $\hat{t}$ exists since all terms above converge to 0. Returning to the differential equation of $\dot{w}_{2,2}$ (Equation (39)):

$$\dot{w}_{2,2}(t) < \epsilon - 2w_{2,2}(t)^2\,. \tag{42}$$

Recalling that $w_{2,2}(t) > 0$ (Lemma 21), it follows that there exists $t_\epsilon \geq \hat{t}$ with $\dot{w}_{2,2}(t_\epsilon) > -\epsilon$ (otherwise $w_{2,2}(t)$ goes to $-\infty$ as $t \to \infty$, in contradiction to the positivity of $w_{2,2}(t)$). For the above t_ϵ, by rearranging the terms in Equation (42) we achieve $w_{2,2}(t_\epsilon) < \sqrt{\epsilon}$. Finally, combined with Equation (41), the result readily follows:

$$\ell(W_{L:1}(t_\epsilon)) = \frac{1}{2}\left[w_{2,2}(t_\epsilon)^2 + 2(w_{1,2}(t_\epsilon) - 1)^2\right] < \epsilon\,,$$

concluding the proof. □

G.7.1 Proof of Lemma 20

The proof proceeds as follows. We initially consider initializations where $W_1(0), \ldots, W_L(0)$ form a *symmetric factorization* of $W_{L:1}(0)$ (Definition 4), and show that this ensures the product matrix stays symmetric. Then, we establish that for every balanced initial factors (Equation (5)) with a positive definite product matrix there exist alternative balanced factors such that: *(i)* the initial product matrix is the same; and *(ii)* the factors form a symmetric factorization of the product matrix. Since the product matrices for the original and the constructed initializations obey the exact same dynamics (Lemma 4), the proof concludes.

Definition 4. We say that the matrices $W_1, W_2, \ldots, W_L \in \mathbb{R}^{d,d}$ form a *symmetric factorization* of $W \in \mathbb{R}^{d,d}$ if $W = W_L W_{L-1} \cdots W_1$ and

$$W_l = W_{L-l+1}^\top \quad , l \in \{1, \ldots, \lfloor L/2 \rfloor + 1\} .$$

A straightforward result is that matrices with a symmetric factorization are symmetric themselves.

Lemma 22. *If a matrix $W \in \mathbb{R}^{d,d}$ has a symmetric factorization, then it is symmetric.*

Proof. Let $W_1, W_2, \ldots, W_L \in \mathbb{R}^{d,d}$ form a symmetric factorization of W. It directly follows that

$$W = W_L W_{L-1} \cdots W_1 = W_1^\top \cdots W_{L-1}^\top W_L^\top = W^\top .$$

□

By Lemma 7, $W_1(t), \ldots, W_L(t), W_{L:1}(t)$ and $\nabla \ell(W_{L:1}(t))$ are analytic, and hence infinitely differentiable, with respect to t. Lemmas 23 and 24 below thus establish that if $W_1(0), \ldots, W_L(0)$ form a symmetric factorization of $W_{L:1}(0)$, then the product matrix stays symmetric for all t.

Lemma 23. *Under the setting of Lemma 20, assume that the matrices $W_1(0), \ldots, W_L(0)$ form a symmetric factorization of $W_{L:1}(0)$ (Definition 4). Then, for all $k \in \mathbb{N} \cup \{0\}$:*

$$W_{L:1}^{(k)}(0) = W_{L:1}^{(k)}(0)^\top , \tag{43}$$

and

$$W_l^{(k)}(0) = W_{L-l+1}^{(k)}(0)^\top \quad , l \in \{1, \ldots, \lfloor L/2 \rfloor + 1\} . \tag{44}$$

Proof. The proof is by induction over k. For $k = 0$, the claim holds directly from the initialization assumption and Lemma 22. For $k \in \mathbb{N}$, suppose the claim is true for all $m \in \mathbb{N} \cup \{0\}$ with $m < k$. We begin by showing Equation (44) holds for k. In turn, this will lead to Equation (43) holding as well. For $l \in [L]$, the dynamics of $W_l(t)$ under gradient flow are

$$W_l^{(1)}(t) = -\tfrac{\partial}{\partial W_l} \phi(W_1(t), W_2(t), \ldots, W_L(t)) = -\prod_{r=l+1}^{L} W_r(t)^\top \cdot G(t) \cdot \prod_{r=1}^{l-1} W_r(t)^\top ,$$

where $G(t) := \nabla \ell(W_{L:1}(t))$ denotes the loss gradient with respect to $W_{L:1}$ at time t. We can explicitly write the k'th ($k \geq 1$) derivative with respect to t of each $W_l(t)$ using the product rule for higher order derivatives:

$$W_l^{(k)}(t) = -\sum_{i_1, \ldots, i_L} \binom{k-1}{i_1, \ldots, i_L} \prod_{r=l+1}^{L} W_r^{(i_r)}(t)^\top \cdot G^{(i_l)}(t) \cdot \prod_{r=1}^{l-1} W_r^{(i_r)}(t)^\top ,$$

where $\sum_{l=1}^{L} i_l = k-1$ and $\binom{k-1}{i_1, \ldots, i_L} = (k-1)! / (i_1! \cdots i_L!)$ for $i_1, \ldots, i_L \in \{0, 1, \ldots, k-1\}$. Taking the transpose of both sides we have:

$$W_l^{(k)}(t)^\top = -\sum_{i_1, \ldots, i_L} \binom{k-1}{i_1, \ldots, i_L} \prod_{1}^{r=l-1} W_r^{(i_r)}(t) \cdot G^{(i_l)}(t)^\top \cdot \prod_{l+1}^{r=L} W_r^{(i_r)}(t) . \tag{45}$$

Turning our attention to $G(t)$, we may write it explicitly as:

$$G(t) = \nabla \ell(W_{L:1}(t)) = \begin{pmatrix} 0 & w_{1,2}(t) - 1 \\ w_{2,1}(t) - 1 & w_{2,2}(t) \end{pmatrix} ,$$

where $\{w_{i,j}(t)\}_{i,j \in [2]}$ are the entries of $W_{L:1}(t)$. For $m < k$, note that when $W_{L:1}^{(m)}(t)$ is symmetric so is $G^{(m)}(t)$. With this in hand, the inductive assumption (Equation (43)) implies that $G^{(m)}(0)$ is symmetric (for all $m < k$). Combined with Equation (44) (for $m < k$, from the inductive assumption), we may write Equation (45) for $t = 0$ as:

$$W_l^{(k)}(0)^\top = -\sum_{i_1, \ldots, i_L} \binom{k-1}{i_1, \ldots, i_L} \prod_{r=L-l+2}^{L} W_r^{(i_{L-r+1})}(0)^\top \cdot G^{(i_l)}(0) \cdot \prod_{r=1}^{L-l} W_r^{(i_{L-r+1})}(0)^\top .$$

Reordering the sum according to $h_r := i_{L-r+1}$ and noticing that $\binom{k-1}{h_1,\ldots,h_L} = \binom{k-1}{i_1,\ldots,i_L}$, we conclude:

$$W_l^{(k)}(0)^\top = -\sum_{h_1,\ldots,h_L} \binom{k-1}{h_1,\ldots,h_L} \prod_{r=L-l+2}^{L} W_r^{(h_r)}(0)^\top \cdot G^{(h_{L-l+1})}(0) \cdot \prod_{r=1}^{L-l} W_r^{(h_r)}(0)^\top .$$

That is,

$$W_l^{(k)}(0)^\top = W_{L-l+1}^{(k)}(0) ,$$

proving Equation (44).

It remains to show that $W_{L:1}^{(k)}(0)$ is symmetric. Similarly to before, we take the k'th derivative of $W_{L:1}(t) := W_L(t) \cdots W_1(t)$ using the product rule:

$$W_{L:1}^{(k)}(t) = \sum_{i_1,\ldots,i_L} \binom{k}{i_1,\ldots,i_L} \prod_{1}^{l=L} W_l^{(i_l)}(t) ,$$

where $\sum_{l=1}^{L} i_l = k$ and $\binom{k}{i_1,\ldots,i_L} = k!/(i_1! \cdots i_L!)$ for $i_1,\ldots,i_L \in \{0,1,\ldots,k\}$. For convenience, we denote $B_{i_1,\ldots,i_L}(t) := \binom{k}{i_1,\ldots,i_L} \prod_1^{l=L} W_l^{(i_l)}(t)$. Pairing up elements in the sum with indices $(i_1,\ldots,i_L)$ that are a reverse order of each other, *i.e.* $(i_1,\ldots,i_L)$ is paired with $(i_L,\ldots,i_1)$:

$$W_{L:1}^{(k)}(t) = \sum_{i_1,\ldots,i_L} \frac{1}{2} \left[B_{i_1,\ldots,i_L}(t)) + B_{i_L,\ldots,i_1}(t) \right] . \tag{46}$$

With Equation (46) in place, we can conclude the proof by showing $W_{L:1}^{(k)}(0)$ is a sum of symmetric matrices. By the inductive assumption for Equation (44), which was established in the first part of the proof for k as well, we have:

$$B_{i_1,\ldots,i_L}(0) = B_{i_L,\ldots,i_1}(0)^\top , \tag{47}$$

for each $(i_1,\ldots,i_L)$. Therefore, the matrix $B_{i_1,\ldots,i_L}(0) + B_{i_L,\ldots,i_1}(0)$ is symmetric. Plugging Equation (47) into Equation (46) with $t = 0$, we arrive at a representation of $W_{L:1}^{(k)}(0)$ as a sum of symmetric matrices. Thus, $W_{L:1}^{(k)}(0)$ is symmetric, completing the proof. □

Lemma 24. *Under the setting of Lemma 20, assume that the matrices $W_1(0),\ldots,W_L(0)$ form a symmetric factorization of $W_{L:1}(0)$ (Definition 4). Then, $W_{L:1}(t)$ is symmetric for all $t \geq 0$.*

Proof. By Lemmas 23 and 10, we may conclude that for all $t \geq 0$:

$$W_l(t) = W_{L-l+1}(t)^\top \quad , l \in \{1,\ldots,\lfloor L/2 \rfloor + 1\} .$$

In words, $W_1(t),\ldots,W_L(t)$ form a symmetric factorization of $W_{L:1}(t)$, and therefore $W_{L:1}(t)$ is symmetric (Lemma 22). □

Going back to the setting of Lemma 20 — initialization is balanced (Equation (5)), but does not necessarily comprise a symmetric factorization — we show that here too the product matrix remains symmetric throughout optimization. To do so, we first construct a factorization of $W_{L:1}(0)$ that is both balanced and symmetric, for which Lemma 24 ensures the product matrix stays symmetric throughout optimization. We then prove that the trajectories of the product matrix for the original and the modified initializations coincide.

Recall that $W_{L:1}(0) = \alpha \cdot I$ and define $\bar{W}_l(0) := \alpha^{\frac{1}{L}} \cdot I$ for $l \in [L]$. It is easily verified that:

- $W_{L:1}(0) = \bar{W}_L(0) \cdots \bar{W}_1(0)$.
- $\bar{W}_l(0) = \bar{W}_{L-l+1}(0)^\top$ for $l \in [L]$.
- $\bar{W}_{l+1}(0)^\top \bar{W}_{l+1}(0) = \bar{W}_l(0) \bar{W}_l(0)^\top$ for $l \in [L-1]$.

Meaning, $\bar{W}_1(0), \ldots, \bar{W}_L(0)$ are balanced, and form a symmetric factorization of $W_{L:1}(0)$. Suppose the factors $\bar{W}_1(t), \ldots, \bar{W}_L(t)$ follow the gradient flow dynamics, with initial values $\bar{W}_1(0), \ldots, \bar{W}_L(0)$, and let $\bar{W}_{L:1}(t) := \bar{W}_L(t) \cdots \bar{W}_1(t)$ be the induced product matrix. From Lemma 24, it follows that $\bar{W}_{L:1}(t)$ is symmetric for all $t \geq 0$.

As characterized in [6] (restated as Lemma 4), if the initial factors are balanced, the product matrix trajectory depends only on its initial value $W_{L:1}(0)$. Since both the original and modified initializations are balanced and have the same product matrix, they lead to the exact same trajectory. Thus, $W_{L:1}(t) = \bar{W}_{L:1}(t)$ for all $t \geq 0$, and specifically, $W_{L:1}(t)$ is symmetric.

The last step is to see that $W_{L:1}(t)$ is not only symmetric, but positive definite as well. Since its initial value $W_{L:1}(0)$ is positive definite, it suffices to show that its eigenvalues do not change sign. By Lemma 6, the determinant of $W_{L:1}(t)$ is positive for all t. Specifically, the product matrix does not have zero eigenvalues. Recalling that $W_{L:1}(t)$ is an analytic function of t (Lemma 7), Theorem 6.1 in [50] implies that its eigenvalues are continuous in t. Therefore, they can not change sign, as that would require them to pass through zero, concluding the proof. □

G.8 Proof of Theorem 2

The proof follows a similar line to that of Theorem 1 (Subappendix G.5), where the differences mostly stem from the fact that the solution set $\widetilde{\mathcal{S}}$ (Equation (14)) is not confined to symmetric matrices, as opposed to the original $\mathcal{S}$ (Equation (7)), slightly complicating the computation of singular values. For the sake of the proof, as mentioned in Subappendix G.1, we omit the time index t when stating results for all $t \geq 0$ or when clear from context. We also let $\{w_{i,j}\}_{i,j \in [2]}$ denote the entries of the product matrix $W_{L:1}$.

We begin by deriving loss-dependent bounds for $|w_{1,1}|$, $\sigma_1(W_{L:1})$ and $\sigma_2(W_{L:1})$. The entries of $W_{L:1}$ can be trivially bounded by the loss as follows:

$$|w_{2,2} - \epsilon| \leq \sqrt{2\ell(W_{L:1})} \quad , \quad |w_{1,2} - z| \leq \sqrt{2\ell(W_{L:1})} \quad , \quad |w_{2,1} - z'| \leq \sqrt{2\ell(W_{L:1})} \, . \tag{48}$$

Lemma 25 below, analogous to Lemma 18 from the proof of Theorem 1, characterizes the relation between $|w_{1,1}|$ and the loss.

Lemma 25. *Suppose $\ell(W_{L:1}) < \min\{z^2/2, z'^2/2\}$. Then:*

$$|w_{1,1}| > \frac{(|z| - \sqrt{2\ell(W_{L:1})})(|z'| - \sqrt{2\ell(W_{L:1})})}{|\epsilon| + \sqrt{2\ell(W_{L:1})}} \geq \frac{|z| \cdot |z'|}{|\epsilon| + \sqrt{2\ell(W_{L:1})}} - (|z| + |z'|) \, .$$

Proof. According to Lemma 6, the determinant of $W_{L:1}$ does not change sign, *i.e.* it remains equal to $\mathrm{sign}(z \cdot z')$ (the initial sign assumed). Under the assumption that $\ell(W_{L:1}) < \min\{z^2/2, z'^2/2\}$, both $w_{1,2}$ and $w_{2,1}$ have the same signs as z and z', respectively, implying that $w_{2,2} \neq 0$ (otherwise we have a contradiction to the sign of the product matrix determinant). If $z \cdot z' > 0$, the determinant is positive as well, and it holds that $w_{1,1}w_{2,2} > w_{1,2}w_{2,1} > 0$. Otherwise, if $z \cdot z' < 0$ we have $w_{1,1}w_{2,2} < w_{1,2}w_{2,1} < 0$. Putting it together we may write $|w_{1,1}w_{2,2}| > |w_{1,2}w_{2,1}|$. Dividing by $|w_{2,2}|$ and applying the bounds from Equation (48) then completes the proof:

$$|w_{1,1}| > \frac{(|z| - \sqrt{2\ell(W_{L:1})})(|z'| - \sqrt{2\ell(W_{L:1})})}{|\epsilon| + \sqrt{2\ell(W_{L:1})}} \geq \frac{|z| \cdot |z'|}{|\epsilon| + \sqrt{2\ell(W_{L:1})}} - (|z| + |z'|) \, .$$

□

We are now able to see that, indeed, the smaller $|\epsilon|$ is compared to $|z \cdot z'|$, the higher $|w_{1,1}|$ will be driven when the loss is minimized. With Lemma 25 in place, we are now able to bound the singular values of $W_{L:1}$.

Lemma 26. *The singular values of $W_{L:1}$ fulfill:*

$$\begin{aligned} \sigma_1(W_{L:1}) &\geq \frac{1}{\sqrt{2}} \cdot |w_{1,1}| - \sqrt{2\ell(W_{L:1})} \, , \\ \sigma_2(W_{L:1}) &\leq 4\,|\epsilon| + \left(4 + \frac{\sqrt{|z| \cdot |z'|}}{\min\{|z|\,, |z'|\}}\right) \sqrt{2\ell(W_{L:1})} \, . \end{aligned} \tag{49}$$

Furthermore, if $\ell(W_{L:1}) < \min\left\{z^2/2 \; , \; z'^2/2\right\}$, the bound on $\sigma_2(W_{L:1})$ may be simplified:

$$\sigma_2(W_{L:1}) \leq 4\,|\epsilon| + 4\sqrt{2\ell(W_{L:1})} \, . \tag{50}$$

Proof. Define $W_{\widetilde{\mathcal{S}}} := \begin{pmatrix} w_{1,1} & z' \\ z & \epsilon \end{pmatrix}$, the orthogonal projection of $W_{L:1}$ onto the solution set $\widetilde{\mathcal{S}}$. From Corollary 8.6.2 in [33] we know that:

$$|\sigma_i(W_{L:1}) - \sigma_i(W_{\widetilde{\mathcal{S}}})| \leq \left\|W_{L:1} - W_{\widetilde{\mathcal{S}}}\right\|_F = \sqrt{2\ell(W_{L:1})} \quad , \; i = 1, 2 \,. \tag{51}$$

This means that any bound on the singular values of $W_{\widetilde{\mathcal{S}}}$ can be transferred to those of $W_{L:1}$ (up to an additive loss-dependent term). It is straightforwardly verified that the squared singular values of $W_{\widetilde{\mathcal{S}}}$ are

$$\begin{aligned} \sigma_1^2(W_{\widetilde{\mathcal{S}}}) &= \frac{1}{2}\left(w_{1,1}^2 + z^2 + z'^2 + \epsilon^2 + \sqrt{\left(w_{1,1}^2 + z^2 + z'^2 + \epsilon^2\right)^2 - 4\left(w_{1,1}\epsilon - zz'\right)^2}\right) \,, \\ \sigma_2^2(W_{\widetilde{\mathcal{S}}}) &= \frac{1}{2}\left(w_{1,1}^2 + z^2 + z'^2 + \epsilon^2 - \sqrt{\left(w_{1,1}^2 + z^2 + z'^2 + \epsilon^2\right)^2 - 4\left(w_{1,1}\epsilon - zz'\right)^2}\right) \,. \end{aligned} \tag{52}$$

Note that the term inside the square roots is non-negative for all $w_{1,1}, z, z', \epsilon$. Since all elements in the expression for $\sigma_1^2(W_{\widetilde{\mathcal{S}}})$ are non-negative, we have $\sigma_1(W_{\widetilde{\mathcal{S}}}) \geq (1/\sqrt{2}) \cdot |w_{1,1}|$. Combining this with Equation (51) completes the lower bound for $\sigma_1(W_{L:1})$.

Next, let $W_{\widetilde{\mathcal{S}}_0} := \begin{pmatrix} w_{1,1} & z' \\ z & 0 \end{pmatrix}$ be the matrix obtained by replacing the bottom-right entry of $W_{\widetilde{\mathcal{S}}}$ by 0. Replacing ϵ with 0 in Equation (52), and applying the identity $a - b = \frac{a^2-b^2}{a+b}$ for $a, b \in \mathbb{R}$ such that $a + b \neq 0$, we get:

$$\begin{aligned} \sigma_2^2(W_{\widetilde{\mathcal{S}}_0}) &= \frac{2z^2z'^2}{w_{1,1}^2 + z^2 + z'^2 + \sqrt{\left(w_{1,1}^2 + z^2 + z'^2\right)^2 - 4z^2z'^2}} \\ &\leq \frac{2z^2z'^2}{w_{1,1}^2 + z^2 + z'^2} \,. \end{aligned} \tag{53}$$

We initially prove Equation (50) holds in the case where $\ell(W_{L:1}) < \min\left\{z^2/2 \,,\; z'^2/2\right\}$. By lifting said assumption we then show that the bound on $\sigma_2(W_{L:1})$ in Equation (49) holds for any loss value. Under the assumption that $\ell(W_{L:1}) < \min\left\{z^2/2 \,,\; z'^2/2\right\}$, taking the square root of both sides in Equation (53), we arrive at the following bound:

$$\begin{aligned} \sigma_2(W_{\widetilde{\mathcal{S}}_0}) &\leq \sqrt{2} \cdot \frac{|z| \cdot |z'|}{\sqrt{w_{1,1}^2 + z^2 + z'^2}} \\ &\leq \sqrt{6} \cdot \frac{|z| \cdot |z'|}{|w_{1,1}| + |z| + |z'|} \\ &\leq \sqrt{6} \cdot \frac{|z| \cdot |z'|}{\frac{|z| \cdot |z'|}{|\epsilon| + \sqrt{2\ell(W_{L:1})}}} \\ &\leq 3\left(|\epsilon| + \sqrt{2\ell(W_{L:1})}\right) \,, \end{aligned}$$

where in the second transition we applied the inequality $\sqrt{w_{1,1}^2 + z^2 + z'^2} \geq (|w_{1,1}| + |z| + |z'|)/\sqrt{3}$, and in the third made use of the bound on $|w_{1,1}|$ (Lemma 25). Applying Corollary 8.6.2 from [33] twice, once for the matrices $W_{L:1}$ and $W_{\widetilde{\mathcal{S}}}$, and another for $W_{\widetilde{\mathcal{S}}}$ and $W_{\widetilde{\mathcal{S}}_0}$, we have:

$$\sigma_2(W_{L:1}) \leq 3\left(|\epsilon| + \sqrt{2\ell(W_{L:1})}\right) + |\epsilon| + \sqrt{2\ell(W_{L:1})} = 4\left(|\epsilon| + \sqrt{2\ell(W_{L:1})}\right) \,,$$

achieving the desired result from Equation (50). It remains to see that the bound on $\sigma_2(W_{L:1})$ in Equation (49) holds regardless of the loss value. When $\ell(W_{L:1}) < \min\left\{z^2/2 \,,\; z'^2/2\right\}$ it obviously holds since it is only looser than the bound already obtained under this assumption. Otherwise, going back to Equation (53), it can be seen that

$$\sigma_2^2(W_{\widetilde{\mathcal{S}}_0}) \leq \frac{2z^2z'^2}{(z - z')^2 + 2\,|z| \cdot |z'|} \leq |z| \cdot |z'| \,.$$

Thus, $\sigma_2(W_{\widetilde{\mathcal{S}}_0}) \leq \sqrt{|z| \cdot |z'|}$. Following the same procedure as before (applying Corollary 8.6.2 from [33]), combined with the fact that $\ell(W_{L:1}) \geq \min\left\{z^2/2 \, , \, z'^2/2\right\}$ concludes the proof:

$$\begin{aligned}
\sigma_2(W_{L:1}) &\leq \sqrt{|z| \cdot |z'|} + |\epsilon| + \sqrt{2\ell(W_{L:1})} \\
&\leq \frac{\sqrt{|z| \cdot |z'|}}{\min\{|z| \, , |z'|\}} \cdot \sqrt{2\ell(W_{L:1})} + |\epsilon| + \sqrt{2\ell(W_{L:1})} \\
&\leq 4\,|\epsilon| + \left(4 + \frac{\sqrt{|z| \cdot |z'|}}{\min\{|z| \, , |z'|\}}\right)\sqrt{2\ell(W_{L:1})} \, .
\end{aligned}$$

□

G.8.1 Proof of Equation (15) (lower bound for quasi-norm)

Turning our attention to $\|W_{L:1}\|$, following the same steps as in the proof of Theorem 1 (Subappendix G.5.1) will lead to a generalized bound. By the triangle inequality:

$$\|W_{L:1}\| \geq \frac{1}{c_{\|\cdot\|}} \left\|w_{1,1}\mathbf{e}_1\mathbf{e}_1^\top\right\| - \left\|W_{L:1} - w_{1,1}\mathbf{e}_1\mathbf{e}_1^\top\right\| \, , \tag{54}$$

where $c_{\|\cdot\|} \geq 1$ is a constant with which $\|\cdot\|$ satisfies the weakened triangle inequality (see Footnote 2). Let us initially assume that $\ell(W_{L:1}) < \min\{z^2/2, z'^2/2\}$. We later lift this assumption, delivering a bound that holds for all loss values. Invoking Equation (48) we may bound the negative term in Equation (54) as follows:

$$\begin{aligned}
\left\|W_{L:1} - w_{1,1}\mathbf{e}_1\mathbf{e}_1^\top\right\| &\leq c_{\|\cdot\|}|w_{2,2}| \left\|\mathbf{e}_2\mathbf{e}_2^\top\right\| + c^2_{\|\cdot\|}\left(|w_{2,1}| \left\|\mathbf{e}_2\mathbf{e}_1^\top\right\| + |w_{1,2}| \left\|\mathbf{e}_1\mathbf{e}_2^\top\right\|\right) \\
&\leq 3c^2_{\|\cdot\|}\left(\max\{|z| \, , |z'|, |\epsilon|\} + \sqrt{2\ell(W_{L:1})}\right) \max_{\substack{i,j\in\{1,2\} \\ (i,j)\neq(1,1)}} \left\|\mathbf{e}_i\mathbf{e}_j^\top\right\| \\
&\leq 6c^2_{\|\cdot\|} \max\{|z| \, , |z'|, |\epsilon|\} \cdot \max_{\substack{i,j\in\{1,2\} \\ (i,j)\neq(1,1)}} \left\|\mathbf{e}_i\mathbf{e}_j^\top\right\| \, ,
\end{aligned}$$

Returning to Equation (54), applying the inequality above and the bound on $|w_{1,1}|$ (Lemma 25) we have:

$$\begin{aligned}
\|W_{L:1}\| &\geq \frac{\left\|\mathbf{e}_1\mathbf{e}_1^\top\right\|}{c_{\|\cdot\|}}\left(\frac{|z| \cdot |z'|}{|\epsilon| + \sqrt{2\ell(W_{L:1})}} - |z| - |z'|\right) - 6c^2_{\|\cdot\|}\max\{|z| \, , |z'|, |\epsilon|\} \max_{\substack{i,j\in\{1,2\} \\ (i,j)\neq(1,1)}} \left\|\mathbf{e}_i\mathbf{e}_j^\top\right\| \\
&\geq \frac{\left\|\mathbf{e}_1\mathbf{e}_1^\top\right\|}{c_{\|\cdot\|}} \cdot \frac{|z| \cdot |z'|}{|\epsilon| + \sqrt{2\ell(W_{L:1})}} - 8c^2_{\|\cdot\|}\max\{|z| \, , |z'|, |\epsilon|\} \cdot \max_{i,j\in\{1,2\}} \left\|\mathbf{e}_i\mathbf{e}_j^\top\right\| \, .
\end{aligned}$$

Since $\|W_{L:1}\|$ is trivially lower bounded by zero, defining the constants

$$a_{\|\cdot\|} := \frac{\left\|\mathbf{e}_1\mathbf{e}_1^\top\right\|}{c_{\|\cdot\|}} \, , \; b_{\|\cdot\|} := \max\left\{\frac{a_{\|\cdot\|} \cdot |z| \cdot |z'|}{|\epsilon| + \min\{|z| \, , |z'|\}}, 8c^2_{\|\cdot\|}\max\{|z| \, , |z'|, |\epsilon|\} \max_{i,j\in\{1,2\}} \left\|\mathbf{e}_i\mathbf{e}_j^\top\right\|\right\} ,$$

allows us, on the one hand, to arrive at a bound of the form:

$$\|W_{L:1}\| \geq a_{\|\cdot\|} \cdot \frac{|z| \cdot |z'|}{|\epsilon| + \sqrt{2\ell(W_{L:1})}} - b_{\|\cdot\|} \, ,$$

and on the other hand, to remove the previous assumption on the loss: in the case where $\ell(W_{L:1}) \geq \min\{z^2/2, z'^2/2\}$, the bound is non-positive and trivially holds. Noticing this is exactly Equation (15) (recall we omitted the time index t), concludes this part of the proof.

G.8.2 Proof of Equation (16) (upper bound for effective rank)

Derivation of the upper bound for effective rank (Definition 1) is initially done under the assumption that $\ell(W_{L:1}) < \min\{z^2/8, z'^2/8\}$. We then remove this assumption, establishing a bound that holds for all loss values.

The bounds on $\sigma_1(W_{L:1})$ and $\sigma_2(W_{L:1})$ in Lemma 26 give:

$$\begin{aligned}
\rho_1(W_{L:1}) &= \frac{\sigma_1(W_{L:1})}{\sigma_1(W_{L:1}) + \sigma_2(W_{L:1})} \\
&\geq \frac{\sigma_1(W_{L:1})}{\sigma_1(W_{L:1}) + 4\left(|\epsilon| + \sqrt{2\ell(W_{L:1})}\right)} \\
&= 1 - \frac{4\left(|\epsilon| + \sqrt{2\ell(W_{L:1})}\right)}{\sigma_1(W_{L:1}) + 4\left(|\epsilon| + \sqrt{2\ell(W_{L:1})}\right)} \\
&\geq 1 - \frac{4\left(|\epsilon| + \sqrt{2\ell(W_{L:1})}\right)}{\frac{1}{\sqrt{2}} \cdot |w_{1,1}| + 4\,|\epsilon| + 3\sqrt{2\ell(W_{L:1})}} \\
&\geq 1 - \frac{4\sqrt{2}\left(|\epsilon| + \sqrt{2\ell(W_{L:1})}\right)}{|w_{1,1}|} \,.
\end{aligned}$$

Additionally, under our assumption that $\ell(W_{L:1}) < \min\{z^2/8, z'^2/8\}$, the bound on $|w_{1,1}|$ in Lemma 25 can be simplified to:

$$|w_{1,1}| \geq \frac{(|z| - \sqrt{2\ell(W_{L:1})})(|z'| - \sqrt{2\ell(W_{L:1})})}{|\epsilon| + \sqrt{2\ell(W_{L:1})}} \geq \frac{\min\{|z|\,, |z'|\}^2}{4\left(|\epsilon| + \sqrt{2\ell(W_{L:1})}\right)} \,.$$

Combining the last two inequalities we have:

$$\rho_2(W_{L:1}) = 1 - \rho_1(W_{L:1}) \leq \frac{16\sqrt{2}\left(|\epsilon| + \sqrt{2\ell(W_{L:1})}\right)^2}{\min\{|z|\,, |z'|\}^2} \,.$$

It is now possible to see that, in accordance with Subsection 3.4, the smaller $|\epsilon|$ is compared to $\min\{|z|\,, |z'|\}$, the closer to zero $\rho_2(W_{L:1})$ becomes as the loss is minimized. Let $h\left(\rho_2(W_{L:1})\right) := -\rho_2(W_{L:1}) \cdot \ln\left(\rho_2(W_{L:1})\right) - (1 - \rho_2(W_{L:1})) \cdot \ln\left(1 - \rho_2(W_{L:1})\right)$ denote the binary entropy function, and recall that the effective rank of the product matrix defined to be $\mathrm{erank}(W_{L:1}) := \exp\{h\left(\rho_2(W_{L:1})\right)\}$. As in the proof of Theorem 1 (Subappendix G.5.2), we may bound the exponent on the interval $[0, \ln(2)]$ by the linear function intersecting it at these points. That is,

$$\mathrm{erank}(W_{L:1}) \leq 1 + \frac{1}{\ln(2)} \cdot h\left(\rho_2(W_{L:1})\right) \,.$$

From Lemma 9 it holds that $h\left(\rho_2(W_{L:1})\right) \leq 2\sqrt{\rho_2(W_{L:1})}$. Plugging this into the inequality above leads to:

$$\begin{aligned}
\mathrm{erank}(W_{L:1}) &\leq 1 + \frac{8 \cdot 2^{\frac{1}{4}}}{\ln(2) \cdot \min\{|z|\,, |z'|\}} \cdot \left(|\epsilon| + \sqrt{2\ell(W_{L:1})}\right) \\
&\leq 1 + \frac{16}{\min\{|z|\,, |z'|\}} \cdot \left(|\epsilon| + \sqrt{2\ell(W_{L:1})}\right) \,,
\end{aligned}$$

where the second transition is a slight simplification of the constants ($2^{1/4}/\ln(2) < 2$). As will be shown below, $\inf_{W' \in \widetilde{\mathcal{S}}} \mathrm{erank}(W') = 1$. We may thus conclude:

$$\mathrm{erank}(W_{L:1}) \leq \inf_{W' \in \widetilde{\mathcal{S}}} \mathrm{erank}(W') + \frac{16}{\min\{|z|\,, |z'|\}} \cdot \left(|\epsilon| + \sqrt{2\ell(W_{L:1})}\right) \,.$$

Notice that when $\ell(W_{L:1}) \geq \min\{z^2/8, z'^2/8\}$ the inequality trivially holds since the right-hand side is greater than 2 (the maximal effective rank for a 2×2 matrix). This establishes Equation (16) (time index is omitted).

It remains to prove that $\inf_{W' \in \widetilde{\mathcal{S}}} \mathrm{erank}(W') = 1$. If $\epsilon \neq 0$, it is trivial since there exists $W' \in \widetilde{\mathcal{S}}$ with $\mathrm{rank}(W') = 1$, meaning $\sigma_2(W') = 0$ and $\mathrm{erank}(W') = 1$. If $\epsilon = 0$, examining the squared singular values of $W' \in \widetilde{\mathcal{S}}$ (Equation (52) with $(W')_{1,1}$ in place of $w_{1,1}$) reveals that $\lim_{(W')_{1,1} \to \infty} \sigma_2(W') = 0$, while $\lim_{(W')_{1,1} \to \infty} \sigma_1(W') = \infty$. Thus, there exists a matrix in $\widetilde{\mathcal{S}}$ with effective rank arbitrarily close to 1. Since the effective rank of any matrix is at least 1, this implies that $\inf_{W' \in \widetilde{\mathcal{S}}} \mathrm{erank}(W') = 1$.

G.8.3 Proof of Equation (17) (upper bound for distance from infimal rank)

We claim that the infimal rank (Definition 2) of $\widetilde{\mathcal{S}}$ is 1. Since $z, z' \neq 0$, it cannot be 0. If $\epsilon \neq 0$, our claim is trivial since there exists $W' \in \widetilde{\mathcal{S}}$ with $\mathrm{rank}(W') = 1$. Otherwise, inspecting the squared singular values of a matrix $W' \in \widetilde{\mathcal{S}}$ (Equation (52) with $(W')_{1,1}$ in place of $w_{1,1}$), we can see that, when $\epsilon = 0$, taking $(W')_{1,1}$ to infinity drives the minimal singular value towards zero ($\lim_{(W')_{1,1} \to \infty} \sigma_2(W') = 0$). Hence, the distance of $\widetilde{\mathcal{S}}$ from the set of matrices with rank 1 or less is 0 in this case as well.

The distance of the product matrix from the infimal rank of $\widetilde{\mathcal{S}}$ is therefore $D(W_{L:1}(t), \mathcal{M}_1) = \sigma_2(W_{L:1}(t))$. From Lemma 26 we have

$$D(W_{L:1}(t), \mathcal{M}_1) \leq 4\,|\epsilon| + \left(4 + \frac{\sqrt{|z| \cdot |z'|}}{\min\{|z|\,, |z'|\}} \right) \sqrt{2\ell(t)}\,,$$

for all $t \geq 0$.

G.8.4 Robustness to change in observed locations

Lastly, we prove that the established bounds (Equations (15), (16) and (17)) are robust to a change in observed locations. Let $(i,j) \in [2] \times [2]$ be the unobserved entry's location. Following proof steps analogous to those in Lemmas 25 and 26 — while recalling our assumption of $\det(W_{L:1}(0))$ having same sign as $z \cdot z'$ if $i = j$ and opposite sign otherwise — yields identical bounds on the unobserved entry and singular values of $W_{L:1}$. Since the derivations of Equations (15), (16) and (17) in Subappendices G.8.1, G.8.2 and G.8.3, respectively, rely solely on the aforementioned bounds, the proof concludes. □

G.9 Proof of Lemma 1

For $l \in [L]$, let $W_l = U_l \Sigma_l V_l^\top$ be a singular value decomposition of W_l, *i.e.* $U_l, V_l \in \mathbb{R}^{d,d}$ are orthogonal matrices, and $\Sigma_l \in \mathbb{R}^{d,d}_{\geq 0}$ is diagonal holding the singular values of W_l in non-increasing order. Define $\{W'_l \in \mathbb{R}^{d,d}\}_{l=1}^L$ by $W'_1 := W_1$, and:

$$W'_l := \prod_{2}^{r=l} \left[U_r V_r^\top\right] \cdot U_1 \Sigma_1 U_1^\top \cdot \prod_{r=2}^{l-1} \left[V_r U_r^\top\right] \quad , \, l = 2, 3, \ldots, L\,.$$

First, for $l \in [L-1]$:

$$\begin{aligned} W'^\top_{l+1} W'_{l+1} &= \prod_{2}^{r=l} \left[U_r V_r^\top\right] U_1 \Sigma_1 U_1^\top \prod_{r=2}^{l+1} \left[V_r U_r^\top\right] \cdot \prod_{2}^{r=l+1} \left[U_r V_r^\top\right] U_1 \Sigma_1 U_1^\top \prod_{r=2}^{l} \left[V_r U_r^\top\right] \\ &= \prod_{2}^{r=l} \left[U_r V_r^\top\right] U_1 \Sigma_1 U_1^\top \prod_{r=2}^{l-1} \left[V_r U_r^\top\right] \cdot \prod_{2}^{r=l-1} \left[U_r V_r^\top\right] U_1 \Sigma_1 U_1^\top \prod_{r=2}^{l} \left[V_r U_r^\top\right] \\ &= W'_l W'^\top_l\,, \end{aligned}$$

i.e. $W'_1, W'_2, \ldots, W'_L$ are balanced.

Second, by induction over $l \in [L]$, we prove that $\|W_l - W'_l\|_F \leq (l-1) \cdot \sqrt{\epsilon}$. For $l = 1$ this is trivial, as $W'_1 = W_1$ by definition. Assume that the bound holds for all $j < l$. Expressing W_l and W'_l in terms of $\{U_r, V_r, \Sigma_r\}_{r=1}^l$ yields:

$$\|W_l - W'_l\|_F = \left\| U_l \Sigma_l V_l^\top - \prod_{2}^{r=l} \left[U_r V_r^\top\right] \cdot U_1 \Sigma_1 U_1^\top \cdot \prod_{r=2}^{l-1} \left[V_r U_r^\top\right] \right\|_F .$$

The Frobenius norm is invariant to multiplication by orthogonal matrices. Thus, we may multiply by $V_l U_l^\top$ from the left:

$$\begin{aligned}\|W_l - W_l'\|_F &= \left\| V_l \Sigma_l V_l^\top - \prod_{2}^{r=l-1} \left[U_r V_r^\top\right] \cdot U_1 \Sigma_1 U_1^\top \cdot \prod_{r=2}^{l-1} \left[V_r U_r^\top\right] \right\|_F \\ &= \left\| \sqrt{W_l^\top W_l} - \sqrt{W_{l-1}' W_{l-1}'^\top} \right\|_F \\ &\leq \left\| \sqrt{W_l^\top W_l} - \sqrt{W_{l-1} W_{l-1}^\top} \right\|_F + \left\| \sqrt{W_{l-1} W_{l-1}^\top} - \sqrt{W_{l-1}' W_{l-1}'^\top} \right\|_F .\end{aligned} \tag{55}$$

Since the unbalancedness magnitude of $W_1, W_2, \ldots, W_L$ is ϵ, from the Powers-Størmer inequality (Lemma 4.1 in [69]) we know that:

$$\left\| \sqrt{W_l^\top W_l} - \sqrt{W_{l-1} W_{l-1}^\top} \right\|_F \leq \sqrt{\|W_l^\top W_l - W_{l-1} W_{l-1}^\top\|_{nuclear}} \leq \sqrt{\epsilon} .$$

Additionally, multiplying by $U_{l-1} V_{l-1}^\top$ from the right:

$$\begin{aligned}\left\| \sqrt{W_{l-1} W_{l-1}^\top} - \sqrt{W_{l-1}' W_{l-1}'^\top} \right\|_F &= \left\| U_{l-1} \Sigma_{l-1} U_{l-1}^\top - \prod_{2}^{r=l-1} \left[U_r V_r^\top\right] U_1 \Sigma_1 U_1^\top \prod_{r=2}^{l-1} \left[V_r U_r^\top\right] \right\|_F \\ &= \left\| U_{l-1} \Sigma_{l-1} V_{l-1}^\top - \prod_{2}^{r=l-1} \left[U_r V_r^\top\right] U_1 \Sigma_1 U_1^\top \prod_{r=2}^{l-2} \left[V_r U_r^\top\right] \right\|_F \\ &= \left\| W_{l-1} - W_{l-1}' \right\|_F \\ &\leq (l-2) \cdot \sqrt{\epsilon} ,\end{aligned}$$

where the last inequality is by the inductive assumption. Going back to Equation (55), we conclude:

$$\|W_l - W_l'\|_F \leq \sqrt{\epsilon} + (l-2) \cdot \sqrt{\epsilon} = (l-1) \cdot \sqrt{\epsilon} .$$

□

G.10 Proof of Lemma 2

Define $g : [0, T] \to \mathbb{R}_{\geq 0}$ by $g(t) := \|\theta(t) - \theta'(t)\|_2^2$. For any $t \in [0, T]$ it holds that:

$$\begin{aligned}\dot{g}(t) &= 2 \left\langle \theta(t) - \theta'(t), \dot{\theta}(t) - \dot{\theta}'(t) \right\rangle \\ &= -2 \left\langle \theta(t) - \theta'(t), \nabla f(\theta(t)) - \nabla f(\theta'(t)) \right\rangle .\end{aligned}$$

By the Cauchy-Schwartz inequality and smoothness of $f(\cdot)$, we have:

$$\begin{aligned}\dot{g}(t) &\leq 2 \|\theta(t) - \theta'(t)\|_2 \cdot \|\nabla f(\theta(t)) - \nabla f(\theta'(t))\|_2 \\ &\leq 2\beta \|\theta(t) - \theta'(t)\|_2^2 \\ &= 2\beta \cdot g(t) .\end{aligned}$$

Let $\bar{t} \in [0, T]$, and suppose that there exists $t_0 \in [0, \bar{t}]$ for which $g(t_0) = 0$. Consider the initial value problem induced by gradient flow over $f(\cdot)$ starting from the point $\theta(t_0) = \theta'(t_0)$. Since its solution on the interval $[t_0, \bar{t}]$ is unique (by the definition of $\theta(\cdot), \theta'(\cdot)$ there exist a solution lying within $\mathcal{D}$, and it not being unique would contradict, *e.g.*, Theorem 2.2 in [81]), it holds that $\theta(t) = \theta'(t)$ for any $t \in [t_0, \bar{t}]$. That is, Equation (19) trivially holds. Now assume that $\forall t \in [0, \bar{t}] : g(t) > 0$. Then:

$$\frac{\dot{g}(t)}{g(t)} \leq 2\beta \implies \int_{t=0}^{\bar{t}} \frac{\dot{g}(t)}{g(t)} dt \leq 2\beta\bar{t} \implies \ln(g(\bar{t})) - \ln(g(0)) \leq 2\beta\bar{t} \implies g(\bar{t}) \leq g(0) \cdot \exp(2\beta\bar{t}) .$$

Taking the square root of both sides in the latter inequality concludes the proof. □

G.11 Proof of Proposition 5

We note that the following proof does not depend on the dimensions of the deep matrix factorization ($d_0, d_1, \ldots, d_L$; see Section 2). That is, Proposition 5 holds for arbitrary dimensions, and is not limited to the square case where $d_0 = d_1 = \ldots = d_L$.

Let $R' > 0$, and define $\mathcal{D}_{R'} := \{(W_1, W_2, \ldots, W_L) : \|(W_1, W_2, \ldots, W_L)\|_F < R'\}$. For any $(W_1, W_2, \ldots, W_L)$ and $(W'_1, W'_2, \ldots, W'_L)$ in $\mathcal{D}_{R'}$:

$$\begin{aligned}
&\|\nabla\phi(W_1, W_2, \ldots, W_L) - \nabla\phi(W'_1, W'_2, \ldots, W'_L)\|_F \\
&= \sqrt{\sum_{l=1}^{L} \left\| \prod_{r=l+1}^{L} W_r^\top \cdot \nabla\ell(W_{L:1}) \cdot \prod_{r=1}^{l-1} W_r^\top - \prod_{r=l+1}^{L} W_r'^\top \cdot \nabla\ell(W'_{L:1}) \cdot \prod_{r=1}^{l-1} W_r'^\top \right\|_F^2} \, . \qquad (56)
\end{aligned}$$

Since $(\nabla\ell(W_{L:1}))_{i,j} = (W_{L:1})_{i,j} - b_{i,j}$ if $(i,j) \in \Omega$, and otherwise $(\nabla\ell(W_{L:1}))_{i,j} = 0$, we have that $\|\nabla\ell(W_{L:1}) - \nabla\ell(W'_{L:1})\|_F \leq \|W_{L:1} - W'_{L:1}\|_F$ and $\|\nabla\ell(W_{L:1})\|_F \leq \|W_{L:1}\|_F + B$. For each $l \in [L]$, Lemma 12 and sub-multiplicativity of the Frobenius norm then yield:

$$\begin{aligned}
&\left\| \prod_{r=l+1}^{L} W_r^\top \cdot \nabla\ell(W_{L:1}) \cdot \prod_{r=1}^{l-1} W_r^\top - \prod_{r=l+1}^{L} W_r'^\top \cdot \nabla\ell(W'_{L:1}) \cdot \prod_{r=1}^{l-1} W_r'^\top \right\|_F \\
&\leq R'^{L-2}(R'^L + B) \cdot \sum_{r=1}^{L} \|W_r - W'_r\|_F + R'^{L-1} \cdot \|\nabla\ell(W_{L:1}) - \nabla\ell(W'_{L:1})\|_F \\
&\leq R'^{L-2}(R'^L + B) \cdot \sum_{r=1}^{L} \|W_r - W'_r\|_F + R'^{L-1} \|W_{L:1} - W'_{L:1}\|_F \\
&\leq R'^{L-2}(R'^L + B) \cdot \sum_{r=1}^{L} \|W_r - W'_r\|_F + R'^{2L-2} \cdot \sum_{r=1}^{L} \|W_r - W'_r\|_F \\
&= \left(2R'^{2L-2} + BR'^{L-2}\right) \cdot \sum_{r=1}^{L} \|W_r - W'_r\|_F \, .
\end{aligned}$$

Plugging the inequality above into Equation (56), we conclude:

$$\begin{aligned}
&\|\nabla\phi(W_1, W_2, \ldots, W_L) - \nabla\phi(W'_1, W'_2, \ldots, W'_L)\|_F \\
&\leq \sqrt{L \cdot (2R'^{2L-2} + BR'^{L-2})^2 \cdot \left(\sum_{l=1}^{L} \|W_l - W'_l\|_F \right)^2} \\
&\leq \sqrt{L^2 \cdot (2R'^{2L-2} + BR'^{L-2})^2 \cdot \sum_{l=1}^{L} \|W_l - W'_l\|_F^2} \\
&= LR'^{L-2}\left(2R'^L + B\right) \cdot \|(W_1, W_2, \ldots, W_L) - (W'_1, W'_2, \ldots, W'_L)\|_F \, ,
\end{aligned}$$

where the second inequality is by $\sum_{l=1}^{L} \|W_l - W'_l\|_F \leq (L \cdot \sum_{l=1}^{L} \|W_l - W'_l\|_F^2)^{0.5}$. That is, the objective $\phi(\cdot)$ is $LR'^{L-2}\left(2R'^L + B\right)$-smooth over $\mathcal{D}_{R'}$. For any $\bar{t} \in [0, T]$, from Lemma 2 we have that:

$$\|\theta(\bar{t}) - \theta'(\bar{t})\|_F \leq \|\theta(0) - \theta'(0)\|_F \cdot \exp\left(LR'^{L-2}\left(2R'^L + B\right) \cdot \bar{t}\right) \, .$$

Notice that R is finite (it is the supremum of a continuous function over a compact domain). Since the inequality above holds for all $R' > R$, taking the limit $R' \to R^+$ yields Equation (20). □

G.12 Proof of Lemma 3

For any $l \in [L]$:

$$\dot{W}_l(t) = -\prod_{r=l+1}^{L} W_r(t)^\top \cdot \nabla\ell(W_{L:1}(t)) \cdot \prod_{r=1}^{l-1} W_r(t)^\top \quad , \forall t \geq 0 \, .$$

This implies that for any $l \in [L-1]$:

$$\dot{W}_l(t) W_l(t)^\top = W_{l+1}(t)^\top \dot{W}_{l+1}(t) \quad , \forall t \geq 0 \, .$$

Adding the transpose of the latter equality to itself, we have:

$$\tfrac{d}{dt}(W_l(t) W_l(t)^\top) = \tfrac{d}{dt}(W_{l+1}(t)^\top W_{l+1}(t)) \quad , \forall t \geq 0 \, .$$

Hence, $W_{l+1}(t)^\top W_{l+1}(t) - W_l(t) W_l(t)^\top = W_{l+1}(0)^\top W_{l+1}(0) - W_l(0) W_l(0)^\top$ for all $t \geq 0$. Taking the nuclear norm of both sides and maximizing over $l \in [L-1]$ yields the desired result. □

G.13 Proof of Theorem 3

The proof of Theorem 1 (Subappendix G.5) relies solely on the fact that $\det(W_{L:1}(t))$ remains positive through time. In particular, it establishes that the bounds on (quasi-)norms, effective rank and distance from infimal rank (Equations (8), (9) and (10) respectively) are guaranteed to hold for any $t \geq 0$ for which $\det(W_{L:1}(t)) > 0$. Therefore, if $\det(W_{L:1}(t)) > 0$ for all $t \geq 0$, the proof concludes. Otherwise, let $T \in [0, \infty)$ be the initial time for which $\det(W_{L:1}(T)) = 0$. Formally, define:

$$T := \inf \left\{ t | t \in [0, \infty) \text{ and } \det(W_{L:}(t)) = 0 \right\} \, .$$

Since $\det(W_{L:1}(t))$ is continuous in t, the set on the right hand side is non-empty, closed, and bounded from below. Thus, T is well defined (*i.e.* $-\infty < T < \infty$), $\det(W_{L:1}(T)) = 0$, and $\det(W_{L:1}(t)) > 0$ for any $t \in [0, T)$ (recall that by assumption the determinant is positive at initialization). That is, the results of Theorem 1 hold over $[0, T)$. Let:

$$R := \begin{cases} \left[\frac{(1-\sqrt{\ell_{init}})^4}{2^{16}} \cdot \ln\left(\frac{1}{\epsilon}\right) \right]^{1/6} - 1 & \text{, if depth } L = 2 \\ \left[\frac{(1-\sqrt{\ell_{init}})^4}{2^{2L+8} L^4} \cdot \frac{1}{\epsilon^{1/32}} \right]^{1/(4L-2)} - 1 & \text{, if depth } L \geq 3 \end{cases} \, . \tag{57}$$

The proof proceeds in two parts. First, assuming that $\max_{l \in [L]} \|W_l(t)\|_F \leq R$ for any $t \in [0, T]$, Subappendix G.13.1 establishes Equation (21) by deriving a lower bound on T. Otherwise, if there exists $t \in [0, T]$ with $\max_{l \in [L]} \|W_l(t)\|_F > R$, Subappendix G.13.2 shows that Equations (22), (23) and (24) jointly hold for some $\bar{t} \in [0, t]$, completing the proof.

G.13.1 Proof of Equation (21) (if weight matrices are bounded by R)

Assume that $\max_{l \in [L]} \|W_l(t)\|_F \leq R$ for any $t \in [0, T]$. The loss during gradient flow is monotonically non-increasing (Lemma 17). This implies that $\ell(W_{L:1}(t)) \leq \ell_{init}$ for all $t \geq 0$. Hence, $\min\{((W_{L:1}(t))_{1,2} - 1)^2, ((W_{L:1}(t))_{2,1} - 1)^2\} \leq \ell_{init}$ and

$$\begin{aligned} \sigma_1(W_{L:1}(t)) &\geq \frac{1}{\sqrt{2}} \|W_{L:1}(t)\|_F \\ &\geq \frac{1}{\sqrt{2}} \max\{(W_{L:1}(t))_{1,2}, (W_{L:1}(t))_{2,1}\} \\ &\geq \frac{1 - \sqrt{\ell_{init}}}{\sqrt{2}} \, , \end{aligned} \tag{58}$$

for all $t \geq 0$. Define:

$$t_0 := \begin{cases} 0 & , \forall t \in [0, T] : \sigma_2(W_{L:1}(t)) < \frac{1-\sqrt{\ell_{init}}}{2} \\ \sup \left\{ t | t \in [0, T] \text{ and } \sigma_2(W_{L:1}(t)) = \frac{1-\sqrt{\ell_{init}}}{2} \right\} & \text{, otherwise} \end{cases} \, .$$

By Lemma 5, $\sigma_2(W_{L:1}(t))$ is a continuous function of t. Combined with the fact that $\sigma_2(W_{L:1}(T)) = 0$ (recall the determinant is zero at T), we have that the set in the alternative case above is non-empty and compact. Thus, t_0 is well defined (*i.e.* $-\infty < t_0 < \infty$). If $t_0 = 0$, then $\sigma_2(W_{L:1}(t_0)) > 0$ since by assumption $\det(W_{L:1}(0)) > 0$. Otherwise, continuity of $\sigma_2(W_{L:1}(t))$ with respect to t implies that $\sigma_2(W_{L:1}(t_0)) = (1 - \sqrt{\ell_{init}})/2 > 0$. Therefore, in either case $t_0 < T$, and $[t_0, T]$ is the interval preceding T over which $\sigma_2(W_{L:1}(t)) \leq (1 - \sqrt{\ell_{init}})/2$.

Fix an arbitrary $\hat{t} \in [t_0, T]$. For conciseness, we hereinafter omit the time index in functions of t at $\hat{t}$, *e.g.* $W_{L:1}$ will be shorthand for $W_{L:1}(\hat{t})$. We now seek to derive a lower bound on $\frac{d}{dt}\sigma_2^2(W_{L:1})$.

Integrating said bound will yield the desired result. By Lemmas 3 and 1, there exist balanced matrices $W_1', W_2', \ldots, W_L'$ satisfying:

$$\|W_l - W_l'\|_F \leq (l-1) \cdot \sqrt{\epsilon} \quad , \forall l \in [L] \,. \tag{59}$$

Applying Lemma 12, and noticing that $\max_{l \in [L]} \|W_l'\|_F \leq R + (L-1) \cdot \sqrt{\epsilon} \leq R + 1$, we bound the distance between the induced product matrices:

$$\|W_{L:1} - W_{L:1}'\|_F \leq (R+1)^{L-1} \cdot \sum_{l=1}^{L} (l-1) \cdot \sqrt{\epsilon} \leq \frac{1}{2} L^2 (R+1)^{L-1} \cdot \sqrt{\epsilon} \,. \tag{60}$$

Consider a gradient flow path originating from $W_1', W_2', \ldots, W_L'$, where for simplicity we regard the initial time of this path as $\hat{t}$. For a balanced flow, Lemma 5 characterizes the movement of the product matrix's singular values.[21] Omitting the time index, by the Cauchy-Schwartz inequality and the fact that the Frobenius norm of an outer product between unit vectors is 1, we have that:

$$\begin{aligned} \left| \frac{d}{dt} \sigma_2(W_{L:1}')^2 \right| &= \left| 2\sigma_2(W_{L:1}') \cdot \frac{d}{dt} \sigma_2(W_{L:1}') \right| \\ &\leq 2L \cdot \sigma_2(W_{L:1}')^{3-2/L} \cdot \|\nabla \ell(W_{L:1}')\|_F \\ &\leq 4L(R+1)^L \cdot \sigma_2(W_{L:1}')^{3-2/L} \,, \end{aligned}$$

where the last transition is by $\|\nabla \ell(W_{L:1}')\|_F \leq \|W_{L:1}'\|_F + \sqrt{2} \leq 2(R+1)^L$. Applying the singular values perturbation bound $\sigma_2(W_{L:1}') \leq \sigma_2(W_{L:1}) + \|W_{L:1} - W_{L:1}'\|_F$ (Corollary 8.6.2 in [33]), the bound in Equation (60), and Jensen's inequality, we arrive at:

$$\begin{aligned} \left| \frac{d}{dt} \sigma_2(W_{L:1}')^2 \right| &\leq 4L(R+1)^L \cdot \left[\sigma_2(W_{L:1}) + \frac{1}{2} L^2 (R+1)^{L-1} \cdot \sqrt{\epsilon} \right]^{3-2/L} \\ &\leq 4 \cdot 2^{2-2/L} L(R+1)^L \cdot \left[\sigma_2(W_{L:1})^{3-2/L} + \left(\frac{1}{2} L^2 (R+1)^{L-1} \cdot \sqrt{\epsilon} \right)^{3-2/L} \right] \\ &\leq 2^4 L(R+1)^L \cdot \sigma_2(W_{L:1})^{3-2/L} + 2 \cdot L^{7-4/L} (R+1)^{4L-5+2/L} \cdot \epsilon^{3/2-1/L} \,. \end{aligned} \tag{61}$$

We turn our attention to $\left| \frac{d}{dt} \sigma_2(W_{L:1})^2 - \frac{d}{dt} \sigma_2(W_{L:1}')^2 \right|$. Recalling that $W_{L:1}$ is a 2-by-2 matrix, its minimal squared singular value can be written as:

$$\sigma_2(W_{L:1})^2 = \frac{1}{2} \left(\|W_{L:1}\|_F^2 - \sqrt{\|W_{L:1}\|_F^4 - 4 \det(W_{L:1})^2} \right) \,.$$

Differentiating with respect to time, while noticing that $(\|W_{L:1}\|_F^4 - 4\det(W_{L:1})^2)^{0.5} = \sigma_1(W_{L:1})^2 - \sigma_2(W_{L:1})^2$, we obtain:

$$\frac{d}{dt} \sigma_2(W_{L:1})^2 = \left\langle W_{L:1}, \frac{d}{dt} W_{L:1} \right\rangle - \frac{\|W_{L:1}\|_F^2 \left\langle W_{L:1}, \frac{d}{dt} W_{L:1} \right\rangle + 2 \det(W_{L:1}) \left\langle A_{W_{L:1}}, \frac{d}{dt} W_{L:1} \right\rangle}{\sigma_1(W_{L:1})^2 - \sigma_2(W_{L:1})^2} \,,$$

with $A_{W_{L:1}} := \begin{pmatrix} (W_{L:1})_{2,2} & -(W_{L:1})_{2,1} \\ -(W_{L:1})_{1,2} & (W_{L:1})_{1,1} \end{pmatrix}$. The same derivation holds for $\frac{d}{dt} \sigma_2(W_{L:1}')$, with $W_{L:1}'$ in place of $W_{L:1}$. Applying the triangle inequality, $\left| \frac{d}{dt} \sigma_2(W_{L:1})^2 - \frac{d}{dt} \sigma_2(W_{L:1}')^2 \right|$ is upper bounded by the sum of the expressions in Equations (62), (63) and (64) below:

$$\left| \left\langle W_{L:1}, \frac{d}{dt} W_{L:1} \right\rangle - \left\langle W_{L:1}', \frac{d}{dt} W_{L:1}' \right\rangle \right| \,, \tag{62}$$

$$\left| \frac{\|W_{L:1}\|_F^2 \left\langle W_{L:1}, \frac{d}{dt} W_{L:1} \right\rangle}{\sigma_1(W_{L:1})^2 - \sigma_2(W_{L:1})^2} - \frac{\|W_{L:1}'\|_F^2 \left\langle W_{L:1}', \frac{d}{dt} W_{L:1}' \right\rangle}{\sigma_1(W_{L:1}')^2 - \sigma_2(W_{L:1}')^2} \right| \,, \tag{63}$$

[21] A technical subtlety is that, since analytic singular values can change order, it is not guaranteed in general that the minimal singular value is analytic in t, *i.e.* the characterization given in Lemma 5 does not necessarily apply to the minimal singular value. However, in our case, over $[t_0, T]$ the maximal singular value is at least $(1 - \sqrt{\ell_{init}})/\sqrt{2}$ (Equation (58)), while the minimal singular value is at most $(1 - \sqrt{\ell_{init}})/2$. Hence, no order change can occur in that time interval, and the movement of the minimal singular is indeed given by Lemma 5.

$$\left| \frac{2\det(W_{L:1}) \left\langle A_{W_{L:1}}, \frac{d}{dt}W_{L:1} \right\rangle}{\sigma_1(W_{L:1})^2 - \sigma_2(W_{L:1})^2} - \frac{2\det(W'_{L:1}) \left\langle A_{W'_{L:1}}, \frac{d}{dt}W'_{L:1} \right\rangle}{\sigma_1(W'_{L:1})^2 - \sigma_2(W'_{L:1})^2} \right| . \tag{64}$$

Lemmas 28, 29 and 30 (with the aid of Lemma (27)) derive upper bounds for the expressions in Equations (62), (63) and (64), respectively. Then, putting it all together, we arrive at:

$$\begin{aligned} \left| \frac{d}{dt}\sigma_2(W_{L:1})^2 - \frac{d}{dt}\sigma_2(W'_{L:1})^2 \right| &\leq 6L^3(R+1)^{4L-3} \cdot \sqrt{\epsilon} + \frac{384L^3(R+1)^{8L-3}}{(1-\sqrt{\ell_{init}})^4} \cdot \sqrt{\epsilon} \\ &\quad + \frac{768L^3(R+1)^{8L-3}}{(1-\sqrt{\ell_{init}})^4} \cdot \sqrt{\epsilon} \\ &\leq \frac{1158L^3(R+1)^{8L-3}}{(1-\sqrt{\ell_{init}})^4} \cdot \sqrt{\epsilon} \,. \end{aligned}$$

Going back to Equation (61), this implies that:

$$\begin{aligned} \left| \frac{d}{dt}\sigma_2(W_{L:1})^2 \right| &\leq \left| \frac{d}{dt}\sigma_2(W'_{L:1})^2 \right| + \left| \frac{d}{dt}\sigma_2(W_{L:1})^2 - \frac{d}{dt}\sigma_2(W'_{L:1})^2 \right| \\ &\leq 2^4 L(R+1)^L \cdot \left(\sigma_2(W_{L:1})^2\right)^{3/2-1/L} + \frac{2^{11}L^7(R+1)^{8L-3}}{(1-\sqrt{\ell_{init}})^4} \cdot \sqrt{\epsilon} \,. \end{aligned}$$

We are now in a position to achieve the sought-after lower bound on T. If $L = 2$, according to Lemma 16:

$$\sigma_2(W_{L:1}(T))^2 \geq \left(\frac{2^7L^6(R+1)^{7L-3}}{(1-\sqrt{\ell_{init}})^4} \cdot \sqrt{\epsilon} + \sigma_{init}^2 \right) \cdot e^{-2^4L(R+1)^L(T-t_0)} - \frac{2^7L^6(R+1)^{7L-3}}{(1-\sqrt{\ell_{init}})^4} \cdot \sqrt{\epsilon} \,,$$

where $\sigma_{init} := \min\{\sigma_2(W_{L:1}(0)), (1-\sqrt{\ell_{init}})/2\} = \sigma_2(W_{L:1}(t_0))$. Due to the fact that T is the initial point in time for which $\det(W_{L:1}(T)) = 0$, the right hand side must be non-positive, in which case:

$$T \geq \frac{1}{2^4L(R+1)^L} \left[\frac{1}{2}\ln\left(\frac{1}{\epsilon}\right) - \ln\left(\frac{2^7L^6(R+1)^{7L-3}}{(1-\sqrt{\ell_{init}})^4\sigma_{init}^2} \right) \right] .$$

Since $\ln\left(2^7L^6(R+1)^{7L-3}/(1-\sqrt{\ell_{init}})^4\sigma_{init}^2\right) \cdot 2^{-4}L^{-1}(R+1)^{-L} \leq \ln\left(e/(1-\sqrt{\ell_{init}})\sigma_{init}\right)$, plugging in the value of R (Equation (57)) leads to the desired result (case $L = 2$ in Equation (21)).

Similarly, for depth $L \geq 3$, Lemma 16 gives the following lower bound:

$$\begin{aligned} &\sigma_2(W_{L:1}(T))^2 \\ &\geq \frac{1}{b_{L,R}^{\frac{2L}{3L-2}} \left[b_{L,R}^{\frac{2L}{3L-2}}(1/2-1/L)(T-t_0) + \left(a_{L,R}^{\frac{2L}{3L-2}} + b_{L,R}^{\frac{2L}{3L-2}} \cdot \sigma_{init}^2 \right)^{\frac{2-L}{2L}} \right]^{\frac{2L}{L-2}}} - \left(\frac{a_{L,R}}{b_{L,R}} \right)^{\frac{2L}{3L-2}} \\ &\geq \frac{1}{b_{L,R}^{\frac{2L}{3L-2}} \left[b_{L,R}^{\frac{2L}{3L-2}} \cdot T + b_{L,R}^{\frac{2-L}{3L-2}} \cdot \sigma_{init}^{\frac{2-L}{L}} \right]^{\frac{2L}{L-2}}} - \left(\frac{a_{L,R}}{b_{L,R}} \right)^{\frac{2L}{3L-2}} , \end{aligned}$$

where $a_{L,R} := \frac{2^{11}L^7(R+1)^{8L-3}}{(1-\sqrt{\ell_{init}})^4} \cdot \sqrt{\epsilon}$ and $b_{L,R} := 2^4L(R+1)^L$. Since $\det(W_{L:1}(T)) = 0$, the lower bound must be non-positive, in which case:

$$T \geq b_{L,R}^{-\frac{2L}{3L-2}} \cdot a_{L,R}^{-\frac{L-2}{3L-2}} - b_{L,R}^{-1} \cdot \sigma_{init}^{-\frac{L-2}{L}} .$$

Noticing that $b_{L,R}^{-1} \leq 2^{-(5L+5)}$, and replacing $a_{L,R}, b_{L,R}$ with their explicit expressions, we arrive at:

$$T \geq 2^{\frac{-19L+22}{3L-2}} (1-\sqrt{\ell_{init}})^{\frac{4L-8}{3L-2}} L^{\frac{-9L+14}{3L-2}} (R+1)^{-\frac{-10L^2+19L-6}{3L-2}} \cdot \epsilon^{-\frac{L-2}{6L-4}} - 2^{-(5L+5)} \sigma_{init}^{-\frac{L-2}{L}} .$$

To simplify the result, we may lower bound the exponent of R by $-4L$. Then, plugging in the value of R (Equation (57)), and applying straightforward bounds to the exponents of 2, $(1-\sqrt{\ell_{init}})$, L and ϵ, concludes this part of the proof (*i.e.* establishes the case $L \geq 3$ in Equation (21)):

$$\begin{aligned}
T &\geq 2^{\frac{4L^2+16L}{2L-1} - \frac{19L-22}{3L-2}} (1-\sqrt{\ell_{init}})^{\frac{4L-8}{3L-2} - \frac{16L}{4L-2}} L^{\frac{16L}{4L-2} - \frac{9L-14}{3L-2}} \cdot \epsilon^{\frac{L}{8(4L-2)} - \frac{L-2}{6L-4}} - 2^{-(5L+5)} \sigma_{init}^{-\frac{L-2}{L}} \\
&\geq 2^{4L/3} (1-\sqrt{\ell_{init}})^{-2} L \cdot \epsilon^{-\frac{3L-8}{32L-16}} - 2^{-(5L+5)} \sigma_{init}^{-\frac{L-2}{L}} \,.
\end{aligned}$$

□

G.13.1.1 Auxiliary Lemmas

Lemma 27. *In the context of the proof for Equation* (21) *(Subappendix G.13.1), the following inequalities hold:*

$$\max\left\{ \left\| \tfrac{d}{dt} W_{L:1} \right\|_F , \left\| \tfrac{d}{dt} W'_{L:1} \right\|_F \right\} \leq 2L(R+1)^{3L-2} \,, \tag{65}$$

$$\left\| \tfrac{d}{dt} W_{L:1} - \tfrac{d}{dt} W'_{L:1} \right\|_F \leq 5L^3 (R+1)^{3L-3} \cdot \sqrt{\epsilon} \,, \tag{66}$$

$$\left| \sigma_1(W_{L:1})^2 - \sigma_2(W_{L:1})^2 - (\sigma_1(W'_{L:1})^2 - \sigma_2(W'_{L:1})^2) \right| \leq 2L^2 (R+1)^{2L-1} \cdot \sqrt{\epsilon} \,, \tag{67}$$

$$\left(\sigma_1(W_{L:1})^2 - \sigma_2(W_{L:1})^2 \right) \left(\sigma_1(W'_{L:1})^2 - \sigma_2(W'_{L:1})^2 \right) \geq (1-\sqrt{\ell_{init}})^4/32 \,. \tag{68}$$

Proof. Starting with Equation (65), the derivative of $W_{L:1}$ with respect to time is:

$$\frac{d}{dt} W_{L:1} = -\sum_{l=1}^{L} \prod_{l+1}^{r=L} W_r \prod_{r=l+1}^{L} W_r^\top \cdot \nabla \ell(W_{L:1}) \cdot \prod_{r=1}^{l-1} W_r^\top \prod_{1}^{r=l-1} W_r \,.$$

Therefore:

$$\begin{aligned}
\left\| \frac{d}{dt} W_{L:1} \right\|_F &\leq \sum_{l=1}^{L} \left\| \prod_{l+1}^{r=L} W_r \prod_{r=l+1}^{L} W_r^\top \cdot \nabla \ell(W_{L:1}) \cdot \prod_{r=1}^{l-1} W_r^\top \prod_{1}^{r=l-1} W_r \right\|_F \\
&\leq 2L(R+1)^{3L-2} \,,
\end{aligned}$$

where the last transition is by sub-multiplicativity of the Frobenius norm, the fact that $\|W_r\|_F \leq R+1$ for $r \in [L]$, and $\|\nabla \ell(W_{L:1})\|_F \leq 2(R+1)^L$. The exact same bound can be derived for $\left\| \frac{d}{dt} W'_{L:1} \right\|_F$, completing the proof of Equation (65).

Moving on to Equation (66):

$$\begin{aligned}
\left\| \frac{d}{dt} W_{L:1} - \frac{d}{dt} W'_{L:1} \right\|_F \leq \sum_{l=1}^{L} \Bigg\| & \prod_{l+1}^{r=L} W_r \prod_{r=l+1}^{L} W_r^\top \cdot \nabla \ell(W_{L:1}) \cdot \prod_{r=1}^{l-1} W_r^\top \prod_{1}^{r=l-1} W_r \\
& - \prod_{l+1}^{r=L} W'_r \prod_{r=l+1}^{L} W_r'^\top \cdot \nabla \ell(W'_{L:1}) \cdot \prod_{r=1}^{l-1} W_r'^\top \prod_{1}^{r=l-1} W'_r \Bigg\|_F \,.
\end{aligned}$$

Noticing that $\|\nabla \ell(W_{L:1}) - \nabla \ell(W'_{L:1})\|_F \leq \|W_{L:1} - W'_{L:1}\|_F$, according to Lemma 12 we may bound each term in the sum by $4(R+1)^{3L-3} \cdot \sum_{l=1}^{L} \|W_l - W'_l\|_F + (R+1)^{2L-2} \cdot \|W_{L:1} - W'_{L:1}\|_F$. Applying the bounds from Equations (59) and (60) then establishes Equation (66):

$$\begin{aligned}
\left\| \frac{d}{dt} W_{L:1} - \frac{d}{dt} W'_{L:1} \right\|_F &\leq L \cdot \left[4L^2 (R+1)^{3L-3} \cdot \sqrt{\epsilon} + \frac{1}{2} L^2 (R+1)^{3L-3} \cdot \sqrt{\epsilon} \right] \\
&\leq 5L^3 (R+1)^{3L-3} \cdot \sqrt{\epsilon} \,.
\end{aligned}$$

Next, Equation (67) is straightforwardly derived using a perturbation bound on the singular values (Corollary 8.6.2 in [33]):

$$\begin{aligned}
\left| \sigma_1(W_{L:1})^2 - \sigma_1(W'_{L:1})^2 \right| &= |\sigma_1(W_{L:1}) - \sigma_1(W'_{L:1})| \cdot |\sigma_1(W_{L:1}) + \sigma_1(W'_{L:1})| \\
&\leq 2(R+1)^L \, \|W_{L:1} - W'_{L:1}\|_F \\
&\leq L^2 (R+1)^{2L-1} \cdot \sqrt{\epsilon} \,,
\end{aligned}$$

where the last transition is by Equation (60). The same derivation shows that a similar inequality holds for $\sigma_2(\cdot)$, establishing Equation (67):

$$\begin{aligned}&\left|\sigma_1(W_{L:1})^2 - \sigma_2(W_{L:1})^2 - (\sigma_1(W'_{L:1})^2 - \sigma_2(W'_{L:1})^2)\right| \\ &\quad \leq \left|\sigma_1(W_{L:1})^2 - \sigma_1(W'_{L:1})^2\right| + \left|\sigma_2(W_{L:1})^2 - \sigma_2(W'_{L:1})^2\right| \\ &\quad \leq 2L^2(R+1)^{2L-1} \cdot \sqrt{\epsilon}\,.\end{aligned}$$

Lastly, recall that $\sigma_2(W_{L:1}(t)) \leq (1 - \sqrt{\ell_{init}})/2$ over $[t_0, T]$. Combined with Equation (58), this implies that $\sigma_1(W_{L:1})^2 - \sigma_2(W_{L:1})^2 \geq (1 - \sqrt{\ell_{init}})^2/4$. Then, by Equation (67) we have:

$$\begin{aligned}&\sigma_1(W'_{L:1})^2 - \sigma_2(W'_{L:1})^2 \\ &\quad \geq \sigma_1(W_{L:1})^2 - \sigma_2(W_{L:1})^2 - \left|\sigma_1(W_{L:1})^2 - \sigma_2(W_{L:1})^2 - (\sigma_1(W'_{L:1})^2 - \sigma_2(W'_{L:1})^2)\right| \\ &\quad \geq (1 - \sqrt{\ell_{init}})^2/4 - 2L^2(R+1)^{2L-1} \cdot \sqrt{\epsilon}\,.\end{aligned}$$

Noticing that $\epsilon \leq \frac{(1-\sqrt{\ell_{init}})^4}{2^{2L+8}L^4(R+1)^{4L-2}}$, in which case $2L^2(R+1)^{2L-1} \cdot \sqrt{\epsilon} \leq (1 - \sqrt{\ell_{init}})^2/8$, concludes the proof:

$$(\sigma_1(W_{L:1})^2 - \sigma_2(W_{L:1})^2)(\sigma_1(W'_{L:1})^2 - \sigma_2(W'_{L:1})^2) \geq \frac{(1 - \sqrt{\ell_{init}})^2}{4} \cdot \frac{(1 - \sqrt{\ell_{init}})^2}{8}\,.$$

□

Lemma 28. *In the context of the proof for Equation* (21) *(Subappendix G.13.1):*

$$\left|\left\langle W_{L:1}, \frac{d}{dt}W_{L:1}\right\rangle - \left\langle W'_{L:1}, \frac{d}{dt}W'_{L:1}\right\rangle\right| \leq 6L^3(R+1)^{4L-3} \cdot \sqrt{\epsilon},$$

and

$$\left|\left\langle A_{W_{L:1}}, \frac{d}{dt}W_{L:1}\right\rangle - \left\langle A_{W'_{L:1}}, \frac{d}{dt}W'_{L:1}\right\rangle\right| \leq 6L^3(R+1)^{4L-3} \cdot \sqrt{\epsilon}.$$

Proof. Starting with the former inequality, adding and subtracting $\langle W'_{L:1}, \frac{d}{dt}W_{L:1}\rangle$, followed by the triangle and Cauchy-Schwartz inequalities, we have that $\left|\langle W_{L:1}, \frac{d}{dt}W_{L:1}\rangle - \langle W'_{L:1}, \frac{d}{dt}W'_{L:1}\rangle\right|$ is bounded by:

$$\|W_{L:1} - W'_{L:1}\|_F \cdot \left\|\frac{d}{dt}W_{L:1}\right\|_F + \|W'_{L:1}\|_F \cdot \left\|\frac{d}{dt}W_{L:1} - \frac{d}{dt}W'_{L:1}\right\|_F\,.$$

From Equation (60) we know that $\|W_{L:1} - W'_{L:1}\|_F \leq (1/2)L^2(R+1)^{L-1} \cdot \sqrt{\epsilon}$. Additionally, by sub-multiplicativity of the Frobenius norm, $\|W'_{L:1}\|_F \leq (R+1)^L$. Applying Equations (65) and (66) from Lemma 27, we conclude:

$$\begin{aligned}\left|\left\langle W_{L:1}, \frac{d}{dt}W_{L:1}\right\rangle - \left\langle W'_{L:1}, \frac{d}{dt}W'_{L:1}\right\rangle\right| &\leq L^3(R+1)^{4L-3} \cdot \sqrt{\epsilon} + 5L^3(R+1)^{4L-3} \cdot \sqrt{\epsilon} \\ &= 6L^3(R+1)^{4L-3} \cdot \sqrt{\epsilon}\,.\end{aligned}$$

For $\left|\langle A_{W_{L:1}}, \frac{d}{dt}W_{L:1}\rangle - \langle A_{W'_{L:1}}, \frac{d}{dt}W'_{L:1}\rangle\right|$, similar proof steps establish the same upper bound since $\left\|A_{W_{L:1}} - A_{W'_{L:1}}\right\|_F = \|W_{L:1} - W'_{L:1}\|_F$ and $\left\|A_{W'_{L:1}}\right\|_F = \|W'_{L:1}\|_F$. □

Lemma 29. *In the context of the proof for Equation* (21) *(Subappendix G.13.1):*

$$\left|\frac{\|W_{L:1}\|_F^2 \left\langle W_{L:1}, \frac{d}{dt}W_{L:1}\right\rangle}{\sigma_1(W_{L:1})^2 - \sigma_2(W_{L:1})^2} - \frac{\|W'_{L:1}\|_F^2 \left\langle W'_{L:1}, \frac{d}{dt}W'_{L:1}\right\rangle}{\sigma_1(W'_{L:1})^2 - \sigma_2(W'_{L:1})^2}\right| \leq \frac{384L^3(R+1)^{8L-3}}{(1 - \sqrt{\ell_{init}})^4} \cdot \sqrt{\epsilon}.$$

Proof. By Equation (68), it suffices to show that:

$$\begin{aligned}&\frac{32}{(1 - \sqrt{\ell_{init}})^4}\left|(\sigma_1(W'_{L:1})^2 - \sigma_2(W'_{L:1})^2)\alpha_{W_{L:1}} - (\sigma_1(W_{L:1})^2 - \sigma_2(W_{L:1})^2)\alpha_{W'_{L:1}}\right| \\ &\quad \leq \frac{384L^3(R+1)^{8L-3}}{(1 - \sqrt{\ell_{init}})^4} \cdot \sqrt{\epsilon}\,,\end{aligned}$$

where $\alpha_{W_{L:1}} := \|W_{L:1}\|_F^2 \left\langle W_{L:1}, \frac{d}{dt}W_{L:1} \right\rangle$, and $\alpha_{W'_{L:1}}$ is defined similarly for $W'_{L:1}$. Focusing on the expression in absolute value, adding and subtracting $(\sigma_1(W_{L:1})^2 - \sigma_2(W_{L:1})^2)\alpha_{W_{L:1}}$, and applying the triangle inequality, leads to:

$$\begin{aligned} \frac{32}{(1-\sqrt{\ell_{init}})^4} \Big[& \left| \sigma_1(W'_{L:1})^2 - \sigma_2(W'_{L:1})^2 - (\sigma_1(W_{L:1})^2 - \sigma_2(W_{L:1})^2) \right| \cdot |\alpha_{W_{L:1}}| \\ & + \left| \sigma_1(W_{L:1})^2 - \sigma_2(W_{L:1})^2 \right| \cdot \left| \alpha_{W_{L:1}} - \alpha_{W'_{L:1}} \right| \Big] \,. \end{aligned} \tag{69}$$

We consider each of these terms separately. From Equation (67) in Lemma 27 we know that $\left| \sigma_1(W'_{L:1})^2 - \sigma_2(W'_{L:1})^2 - (\sigma_1(W_{L:1})^2 - \sigma_2(W_{L:1})^2 \right| \leq 2L^2(R+1)^{2L-1} \cdot \sqrt{\epsilon}$. Equation (65) and the Cauchy-Schwartz inequality give:

$$|\alpha_{W_{L:1}}| \leq \|W_{L:1}\|_F^3 \cdot \left\| \frac{d}{dt} W_{L:1} \right\|_F \leq 2L(R+1)^{6L-2} \,.$$

Additionally, $\left| \sigma_1(W_{L:1})^2 - \sigma_2(W_{L:1})^2 \right| \leq \sigma_1(W_{L:1})^2 + \sigma_2(W_{L:1})^2 = \|W_{L:1}\|_F^2 \leq (R+1)^{2L}$. By adding and subtracting $\|W'_{L:1}\|_F^2 \left\langle W_{L:1}, \frac{d}{dt}W_{L:1} \right\rangle$, we may upper bound $\left| \alpha_{W_{L:1}} - \alpha_{W'_{L:1}} \right|$ by:

$$\begin{aligned} & \left| \|W_{L:1}\|_F^2 - \|W'_{L:1}\|_F^2 \right| \left| \left\langle W_{L:1}, \frac{d}{dt} W_{L:1} \right\rangle \right| + \|W'_{L:1}\|_F^2 \left| \left\langle W_{L:1}, \frac{d}{dt} W_{L:1} \right\rangle - \left\langle W'_{L:1}, \frac{d}{dt} W'_{L:1} \right\rangle \right| \\ & \quad \leq |\|W_{L:1}\|_F + \|W'_{L:1}\|_F| \cdot |\|W_{L:1}\|_F - \|W'_{L:1}\|_F| \cdot \|W_{L:1}\|_F \left\| \frac{d}{dt} W_{L:1} \right\|_F \\ & \qquad + \|W'_{L:1}\|_F^2 \left| \left\langle W_{L:1}, \frac{d}{dt} W_{L:1} \right\rangle - \left\langle W'_{L:1}, \frac{d}{dt} W'_{L:1} \right\rangle \right| \\ & \quad \leq 2(R+1)^{2L} \, |\|W_{L:1}\|_F - \|W'_{L:1}\|_F| \cdot \left\| \frac{d}{dt} W_{L:1} \right\|_F \\ & \qquad + (R+1)^{2L} \left| \left\langle W_{L:1}, \frac{d}{dt} W_{L:1} \right\rangle - \left\langle W'_{L:1}, \frac{d}{dt} W'_{L:1} \right\rangle \right| \\ & \quad \leq 2L^3(R+1)^{6L-3} \cdot \sqrt{\epsilon} + 6L^3(R+1)^{6L-3} \cdot \sqrt{\epsilon} \\ & \quad = 8L^3(R+1)^{6L-3} \cdot \sqrt{\epsilon} \,. \end{aligned}$$

where the third transition is due to Equations (60), (65), and Lemma 28. Putting it all together, the expression in Equation (69) is upper bounded by:

$$\begin{aligned} & \frac{32}{(1-\sqrt{\ell_{init}})^4} \left[4L^3(R+1)^{8L-3} \cdot \sqrt{\epsilon} \cdot + 8L^3(R+1)^{8L-3} \cdot \sqrt{\epsilon} \right] \\ & \quad = \frac{384L^3(R+1)^{8L-3}}{(1-\sqrt{\ell_{init}})^4} \cdot \sqrt{\epsilon} \,. \end{aligned}$$

□

Lemma 30. *In the context of the proof for Equation* (21) *(Subappendix G.13.1):*

$$\left| \frac{2\det(W_{L:1}) \left\langle A_{W_{L:1}}, \frac{d}{dt} W_{L:1} \right\rangle}{\sigma_1(W_{L:1})^2 - \sigma_2(W_{L:1})^2} - \frac{2\det(W'_{L:1}) \left\langle A_{W'_{L:1}}, \frac{d}{dt} W'_{L:1} \right\rangle}{\sigma_1(W'_{L:1})^2 - \sigma_2(W'_{L:1})^2} \right| \leq \frac{768L^3(R+1)^{8L-3}}{(1-\sqrt{\ell_{init}})^4} \cdot \sqrt{\epsilon} \,.$$

Proof. The proof follows a line similar to that of Lemma 29. Applying the bound from Equation (68) in Lemma 27, we arrive at the following upper bound for the left hand side above:

$$\frac{64}{(1-\sqrt{\ell_{init}})^4} \left| (\sigma_1(W'_{L:1})^2 - \sigma_2(W'_{L:1})^2)\beta_{W_{L:1}} - (\sigma_1(W_{L:1})^2 - \sigma_2(W_{L:1})^2)\beta_{W'_{L:1}} \right| \,,$$

where $\beta_{W_{L:1}} := \det(W_{L:1}) \left\langle A_{W_{L:1}}, \frac{d}{dt} W_{L:1} \right\rangle$, and $\beta_{W'_{L:1}}$ is defined similarly for $W'_{L:1}$. Focusing on the expression in absolute value, we add and subtract $(\sigma_1(W_{L:1})^2 - \sigma_2(W_{L:1})^2)\beta_{W_{L:1}}$ and apply the triangle inequality:

$$\begin{aligned} \frac{64}{(1-\sqrt{\ell_{init}})^4} \Big[& \left| \sigma_1(W'_{L:1})^2 - \sigma_2(W'_{L:1})^2 - (\sigma_1(W_{L:1})^2 - \sigma_2(W_{L:1})^2) \right| \cdot |\beta_{W_{L:1}}| \\ & + \left| \sigma_1(W_{L:1})^2 - \sigma_2(W_{L:1})^2 \right| \cdot \left| \beta_{W_{L:1}} - \beta_{W'_{L:1}} \right| \Big] \,. \end{aligned} \tag{70}$$

We treat each of the four terms separately. From Equation (67) in Lemma 27 we know that:

$$\left|\sigma_1(W'_{L:1})^2 - \sigma_2(W'_{L:1})^2 - (\sigma_1(W_{L:1})^2 - \sigma_2(W_{L:1})^2\right| \leq 2L^2(R+1)^{2L-1} \cdot \sqrt{\epsilon}\,.$$

Since $|\det(W_{L:1})| = \sigma_1(W_{L:1}) \cdot \sigma_2(W_{L:1}) \leq \|W_{L:1}\|_F^2$, the Cauchy-Schwartz inequality and Equation (65) from Lemma 27 give:

$$|\beta_{W_{L:1}}| \leq \|W_{L:1}\|_F^3 \left\| \frac{d}{dt} W_{L:1} \right\|_F \leq 2L(R+1)^{6L-2}\,.$$

Furthermore, $\left|\sigma_1(W_{L:1})^2 - \sigma_2(W_{L:1})^2\right| \leq \|W_{L:1}\|_F^2 \leq (R+1)^{2L}$. The remaining term is $\left|\beta_{W_{L:1}} - \beta_{W'_{L:1}}\right|$. Adding and subtracting $\det(W'_{L:1}) \left\langle A_{W_{L:1}}, \frac{d}{dt} W_{L:1} \right\rangle$, it can be upper bounded by:

$$\begin{aligned} &|\det(W_{L:1}) - \det(W'_{L:1})| \left| \left\langle A_{W_{L:1}}, \frac{d}{dt} W_{L:1} \right\rangle \right| \\ &+ |\det(W'_{L:1})| \left| \left\langle A_{W_{L:1}}, \frac{d}{dt} W_{L:1} \right\rangle - \left\langle A_{W'_{L:1}}, \frac{d}{dt} W'_{L:1} \right\rangle \right|\,. \end{aligned}$$

From Lemma 28 we have that $\left|\left\langle A_{W_{L:1}}, \frac{d}{dt} W_{L:1} \right\rangle - \left\langle A_{W'_{L:1}}, \frac{d}{dt} W'_{L:1} \right\rangle\right| \leq 6L^3(R+1)^{4L-3} \cdot \sqrt{\epsilon}$. Furthermore, Theorem 2.12 in [42] implies that:

$$|\det(W_{L:1}) - \det(W'_{L:1})| \leq 2\,\|W_{L:1} - W'_{L:1}\|_F \max\{\|W_{L:1}\|_F\,, \|W'_{L:1}\|_F\}\,.$$

Thus, with the use of Equations (60) and (65), we have:

$$\begin{aligned} \left|\beta_{W_{L:1}} - \beta_{W'_{L:1}}\right| &\leq 2L(R+1)^{4L-2} \cdot |\det(W_{L:1}) - \det(W'_{L:1})| + 6L^3(R+1)^{6L-3} \cdot \sqrt{\epsilon} \\ &\leq 2L^3(R+1)^{6L-3} \cdot \sqrt{\epsilon} + 6L^3(R+1)^{6L-3} \cdot \sqrt{\epsilon} \\ &= 8L^3(R+1)^{6L-3} \cdot \sqrt{\epsilon}\,. \end{aligned}$$

Put together, the expression in Equation (70) is upper bounded by:

$$\begin{aligned} &\frac{64}{(1-\sqrt{\ell_{init}})^4} \left[4L^3(R+1)^{8L-3} \cdot \sqrt{\epsilon} + 8L^3(R+1)^{8L-3} \cdot \sqrt{\epsilon}\right] \\ &= \frac{768}{(1-\sqrt{\ell_{init}})^4} L^3(R+1)^{8L-3} \cdot \sqrt{\epsilon}\,. \end{aligned}$$

□

G.13.2 Proof of Equations (22), (23) and (24) (if weight matrices are not bounded by R)

Assume that there exists $t \in [0,T]$ with $\max_{l\in[L]} \|W_l(t)\|_F > R$. We examine the initial time at which the Frobenius norm of one of the weight matrices reaches R. Formally, define:

$$\bar{t} := \inf\left\{\hat{t} | \hat{t} \in [0,t] \text{ and } \exists l \in [L] \text{ s.t. } \left\|W_l(\hat{t})\right\|_F = R\right\}\,.$$

Since $R \geq \max_{l\in[L]} \|W_l(0)\|_F$ — implied by the assumption on ϵ — and $W_1(t), W_2(t), \ldots, W_L(t)$ are continuous functions of t, the set on the right hand side is non-empty and compact. Hence, $\bar{t}$ is well defined (*i.e.* $-\infty < \bar{t} < \infty$), with $\max_{l\in[L]} \|W_l(\bar{t})\|_F = R$.

Next, we derive a lower bound on $|(W_{L:1}(\bar{t}))_{1,1}|$ — the absolute value of the unobserved entry. According to Lemmas 3 and 1, there exist balanced matrices $W'_1, W'_2, \ldots, W'_L$ satisfying:

$$\|W_l(\bar{t}) - W'_l\|_F \leq (l-1) \cdot \sqrt{\epsilon} \quad , \forall l \in [L]\,. \tag{71}$$

Applying Lemma 12, and noticing that $\max_{l\in[L]} \|W'_l\|_F \leq R + (L-1) \cdot \sqrt{\epsilon} \leq R+1$, we bound the distance between the induced product matrices:

$$\|W_{L:1}(\bar{t}) - W'_{L:1}\|_F \leq (R+1)^{L-1} \cdot \sum_{l=1}^{L} (l-1) \cdot \sqrt{\epsilon} \leq \frac{1}{2} L^2 (R+1)^{L-1} \cdot \sqrt{\epsilon}\,. \tag{72}$$

Since $W_1', W_2', \ldots, W_L'$ are balanced, they have the same singular values. In particular, this means that $\|W_l'\|_F = \|W_{l'}'\|_F$ for any $l, l' \in [L]$. Additionally, by Lemma 8 we know that $\sigma_1(W_{L:1}') = \sigma_1(W_1')^L$. Thus:

$$\begin{aligned}
\|W_{L:1}'\|_F &\geq \sigma_1(W_{L:1}') \\
&= \sigma_1(W_1')^L \\
&\geq \left(\frac{1}{\sqrt{2}}\|W_1'\|_F\right)^L \\
&= 2^{-\frac{L}{2}} \max_{l\in[L]} \|W_l'\|_F^L \\
&\geq 2^{-\frac{L}{2}} \cdot \left(R - (L-1)\cdot\sqrt{\epsilon}\right)^L ,
\end{aligned}$$

where the last transition is by Equation (71) and the fact that $\max_{l\in[L]} \|W_l(\bar{t})\|_F = R$. The inequality above, combined with Equation (72), leads to:

$$\begin{aligned}
\|W_{L:1}(\bar{t})\|_F &\geq \|W_{L:1}'\|_F - \frac{1}{2}L^2 (R+1)^{L-1} \cdot \sqrt{\epsilon} \\
&\geq 2^{-\frac{L}{2}} \cdot \left(R - (L-1)\cdot\sqrt{\epsilon}\right)^L - \frac{1}{2}L^2 (R+1)^{L-1} \cdot \sqrt{\epsilon} .
\end{aligned}$$

From the definition of R it holds that $\epsilon \leq \min\left\{ \left(\frac{(1-1/\sqrt{2})\cdot R}{L-1}\right)^2, \frac{(R/2)^{2L}}{L^4(R+1)^{2L-2}} \right\}$, and therefore:

$$\|W_{L:1}(\bar{t})\|_F \geq \left(\frac{R}{2}\right)^L - \frac{1}{2}\left(\frac{R}{2}\right)^L = \frac{R^L}{2^{L+1}} . \tag{73}$$

Recall that $\ell_{init} < \ell(0) = 1$. The loss during gradient flow is monotonically non-increasing with respect to time (Lemma 17), hence, $\ell(W_{L:1}(\bar{t})) < 1$. This leads to the following bounds on the observed entries:

$$|(W_{L:1}(\bar{t}))_{1,2}| \leq 1+\sqrt{2} \quad , \quad |(W_{L:1}(\bar{t}))_{2,1}| \leq 1+\sqrt{2} \quad , \quad |(W_{L:1}(\bar{t}))_{2,2}| \leq \sqrt{2} . \tag{74}$$

Applying these bounds, we obtain:

$$\|W_{L:1}(\bar{t})\|_F \leq |(W_{L:1}(\bar{t}))_{1,1}| + \sqrt{2(1+\sqrt{2})^2 + 2} \leq |(W_{L:1}(\bar{t}))_{1,1}| + 4 .$$

Combined with Equation (73), we have that:

$$|(W_{L:1}(\bar{t}))_{1,1}| \geq \frac{R^L}{2^{L+1}} - 4 . \tag{75}$$

We are now in a position to establish Equations (22), (23) and (24). Starting with Equation (22), for any quasi-norm $\|\cdot\|$ it holds that:

$$\|W_{L:1}(\bar{t})\| \geq \frac{1}{c_{\|\cdot\|}} \left\|(W_{L:1}(\bar{t}))_{1,1}\mathbf{e}_1\mathbf{e}_1^\top\right\| - \left\|W_{L:1}(\bar{t}) - (W_{L:1}(\bar{t}))_{1,1}\mathbf{e}_1\mathbf{e}_1^\top\right\| , \tag{76}$$

where $c_{\|\cdot\|} \geq 1$ is a constant for which $\|\cdot\|$ satisfies the weakened triangle inequality (see Footnote 2). Subsequent applications of the weakened triangle inequality, together with homogeneity of $\|\cdot\|$ and the bounds on the observed entries (Equation (74)), give:

$$\left\|W_{L:1}(\bar{t}) - (W_{L:1}(\bar{t}))_{1,1}\mathbf{e}_1\mathbf{e}_1^\top\right\| \leq 8c_{\|\cdot\|}^2 \max_{i,j\in\{1,2\}} \left\|\mathbf{e}_i\mathbf{e}_j^\top\right\| .$$

Plugging the inequality above into Equation (76), and lower bounding $|(W_{L:1}(\bar{t}))_{1,1}|$ according to Equation (75), we have:

$$\begin{aligned}
\|W_{L:1}(\bar{t})\| &\geq \frac{\left\|\mathbf{e}_1\mathbf{e}_1^\top\right\|}{c_{\|\cdot\|}} \cdot |(W_{L:1}(\bar{t}))_{1,1}| - 8c_{\|\cdot\|}^2 \max_{i,j\in\{1,2\}} \left\|\mathbf{e}_i\mathbf{e}_j^\top\right\| \\
&\geq \frac{\left\|\mathbf{e}_1\mathbf{e}_1^\top\right\|}{2^{L+1}c_{\|\cdot\|}} \cdot R^L - 12c_{\|\cdot\|}^2 \max_{i,j\in\{1,2\}} \left\|\mathbf{e}_i\mathbf{e}_j^\top\right\| .
\end{aligned} \tag{77}$$

The assumption on the size of ϵ implies that $R \geq 32$. Hence:

$$R \geq \begin{cases} \frac{1}{2}\left[\frac{(1-\sqrt{\ell_{init}})^4}{2^{16}} \cdot \ln\left(\frac{1}{\epsilon}\right)\right]^{1/6} & \text{, if depth } L = 2 \\ \frac{1}{2}\left[\frac{(1-\sqrt{\ell_{init}})^4}{2^{2L+8}L^4} \cdot \frac{1}{\epsilon^{1/32}}\right]^{1/(4L-2)} & \text{, if depth } L \geq 3 \end{cases} . \tag{78}$$

Applying Equation (78) to Equation (77), we obtain:

$$\|W_{L:1}(\bar{t})\| \geq \begin{cases} \frac{\left\|\mathbf{e}_1\mathbf{e}_1^\top\right\|(1-\sqrt{\ell_{init}})^{4/3}}{2^{31/3}c_{\|\cdot\|}} \cdot \ln\left(\frac{1}{\epsilon}\right)^{1/3} - 12c_{\|\cdot\|}^2 \max\limits_{i,j\in\{1,2\}} \left\|\mathbf{e}_i\mathbf{e}_j^\top\right\| & \text{, if depth } L = 2 \\ \frac{\left\|\mathbf{e}_1\mathbf{e}_1^\top\right\|(1-\sqrt{\ell_{init}})^{\frac{2L}{2L-1}}}{2^{\frac{5L^2+4L-1}{2L-1}}L^{\frac{2L}{2L-1}}c_{\|\cdot\|}} \cdot \epsilon^{-\frac{L}{128L-64}} - 12c_{\|\cdot\|}^2 \max\limits_{i,j\in\{1,2\}} \left\|\mathbf{e}_i\mathbf{e}_j^\top\right\| & \text{, if depth } L \geq 3 \end{cases} .$$

For clarity, we simplify the exponents in the lower bound above. In the case of $L = 2$, we use the fact that $2^{31/3} \leq 2^{11}$. For $L \geq 3$, noticing that $2L/(2L-1) \leq 6/5$ and $(5L^2+4L-1)/(2L-1) \leq 4L$ completes the proof of Equation (22).

Next, we derive the upper bound for effective rank (Equation (23)). Let $h(p) := -p \cdot \ln(p) - (1-p) \cdot \ln(1-p)$ be the binary entropy function, defined over $[0,1]$. Recall that the effective rank of $W_{L:1}(\bar{t})$ is defined to be $\mathrm{erank}(W_{L:1}(\bar{t})) := \exp\{h(\rho_2(W_{L:1}(\bar{t})))\}$, where $\rho_2(W_{L:1}(\bar{t})) := \sigma_2(W_{L:1}(\bar{t}))/(\sigma_1(W_{L:1}(\bar{t})) + \sigma_2(W_{L:1}(\bar{t})))$. Since the exponent function is convex, it is upper bounded on the interval $[0, \ln(2)]$ by the linear function that intersects it at these points. That is:

$$\mathrm{erank}(W_{L:1}(\bar{t})) \leq 1 + \frac{1}{\ln(2)} \cdot h(\rho_2(W_{L:1}(\bar{t}))) \ .$$

From Lemma 9 we know that $h(\rho_2(W_{L:1}(\bar{t}))) \leq 2\sqrt{\rho_2(W_{L:1}(\bar{t}))}$. Combined with the fact that $\inf_{W'\in\mathcal{S}} \mathrm{erank}(W') = 1$ (Proposition 2), this leads to the following upper bound:

$$\mathrm{erank}(W_{L:1}(\bar{t})) \leq \inf_{W'\in\mathcal{S}} \mathrm{erank}(W') + \frac{2}{\ln(2)} \cdot \sqrt{\rho_2(W_{L:1}(\bar{t}))} \, . \tag{79}$$

We now lower bound $\rho_1(W_{L:1}(\bar{t})) := \sigma_1(W_{L:1}(\bar{t}))/(\sigma_1(W_{L:1}(\bar{t})) + \sigma_2(W_{L:1}(\bar{t})))$ as follows:

$$\begin{aligned} \rho_1(W_{L:1}(\bar{t})) &\geq \frac{\sigma_1(W_{L:1})}{\sigma_1(W_{L:1}) + 2^{L+2}/R^L + \sqrt{2\ell(W_{L:1}(\bar{t}))}} \\ &= 1 - \frac{2^{L+2}/R^L + \sqrt{2\ell(W_{L:1}(\bar{t}))}}{\sigma_1(W_{L:1}) + 2^{L+2}/R^L + \sqrt{2\ell(W_{L:1}(\bar{t}))}} \\ &\geq 1 - \frac{2^{L+2}/R^L + \sqrt{2\ell(W_{L:1}(\bar{t}))}}{R^L/2^{L+2} + 2^{L+2}/R^L} \\ &\geq 1 - \frac{2^{L+2}/R^L + \sqrt{2\ell(W_{L:1}(\bar{t}))}}{R^L/2^{L+2}} \\ &= 1 - \left(\frac{2^{L+2}}{R^L}\right)^2 - \sqrt{2\ell(W_{L:1}(\bar{t}))} \cdot \frac{2^{L+2}}{R^L} \, , \end{aligned}$$

where the first and second inequalities are due to Lemma 31. Since $2^{L+2}/R^L < 1$ and $\ell(W_{L:1}(\bar{t})) < 1$, it holds that $\rho_1(W_{L:1}(\bar{t})) \geq 1 - 2^{L+4}/R^L$, or equivalently, $\rho_2(W_{L:1}(\bar{t})) \leq 2^{L+4}/R^L$. Going back to Equation 79, we have:

$$\mathrm{erank}(W_{L:1}(\bar{t})) \leq \inf_{W'\in\mathcal{S}} \mathrm{erank}(W') + \frac{2^{L/2+3}}{\ln(2) \cdot R^{L/2}} \, .$$

Applying the lower bound on R from Equation (78), we arrive at:

$$\mathrm{erank}(W_{L:1}(\bar{t})) \leq \begin{cases} \inf_{W'\in\mathcal{S}} \mathrm{erank}(W') + \frac{2^{23/3}}{\ln(2)(1-\sqrt{\ell_{init}})^{2/3}} \cdot \ln\left(\frac{1}{\epsilon}\right)^{-1/6} & \text{, if depth } L = 2 \\ \inf_{W'\in\mathcal{S}} \mathrm{erank}(W') + \frac{2^{\frac{5L^2+14L-6}{4L-2}}L^{\frac{L}{2L-1}}}{\ln(2)(1-\sqrt{\ell_{init}})^{\frac{L}{2L-1}}} \cdot \epsilon^{\frac{L}{256L-128}} & \text{, if depth } L \geq 3 \end{cases} .$$

In the case of $L = 2$, we simplify the upper bound using the fact that $2^{23/3}/\ln(2) \leq 2^9$. For $L \geq 3$, noticing that $L/(2L-1) \leq 1$ and $(5L^2 + 14L - 6)/(4L-2) \leq 2L + 4$ establishes Equation (23).

Finally, we turn our attention to the upper bound for distance from infimal rank (Equation (24)). By Proposition 2, the infimal rank of $\mathcal{S}$ is 1. Therefore, Lemma 31 yields:

$$D(W_{L:1}(\bar{t}), \mathcal{M}_1) = \sigma_2(W_{L:1}(\bar{t})) \leq \frac{2^{L+2}}{R^L} + \sqrt{2\ell(W_{L:1}(\bar{t}))}\,.$$

Applying the lower bound on R from Equation (78), we arrive at:

$$D(W_{L:1}(\bar{t}), \mathcal{M}_{\mathrm{irank}(\mathcal{S})}) \leq \begin{cases} \frac{2^{34/3}}{(1-\sqrt{\ell_{init}})^{4/3}} \cdot \ln\left(\frac{1}{\epsilon}\right)^{-1/3} + \sqrt{2\ell(W_{L:1}(\bar{t}))} & \text{, if depth } L = 2 \\ \frac{2^{\frac{5L^2+6L-2}{2L-1}} L^{\frac{2L}{2L-1}}}{(1-\sqrt{\ell_{init}})^{\frac{2L}{2L-1}}} \cdot \epsilon^{\frac{L}{128L-64}} + \sqrt{2\ell(W_{L:1}(\bar{t}))} & \text{, if depth } L \geq 3 \end{cases}.$$

In the case of $L = 2$, we simplify the upper bound using the fact that $2^{34/3} \leq 2^{12}$. For $L \geq 3$, noticing that $2L/(2L-1) \leq 6/5$ and $(5L^2 + 6L - 2)/(2L-1) \leq 3L + 4$ concludes the proof of Equation (24). □

G.13.2.1 Auxiliary Lemma

Lemma 31. *In the context of the proof for Equations* (22)*,* (23) *and* (24) *(Subappendix G.13.2), the singular values of $W_{L:1}(\bar{t})$ fulfill:*

$$\sigma_1(W_{L:1}(\bar{t})) \geq \frac{R^L}{2^{L+2}} - \sqrt{2\ell(W_{L:1}(\bar{t}))} \quad , \quad \sigma_2(W_{L:1}(\bar{t})) \leq \frac{2^{L+2}}{R^L} + \sqrt{2\ell(W_{L:1}(\bar{t}))}\,. \tag{80}$$

Proof. Define $W_{\mathcal{S}} := \begin{pmatrix} (W_{L:1}(\bar{t}))_{1,1} & 1 \\ 1 & 0 \end{pmatrix}$, the orthogonal projection of $W_{L:1}(\bar{t})$ onto the solution set $\mathcal{S}$ (Equation (7)). Suppose that $(W_{L:1}(\bar{t}))_{1,1} \geq 0$. The largest singular value of $W_{\mathcal{S}}$ can be written as:

$$\sigma_1(W_{\mathcal{S}}) = \max_{i=1,2} |\lambda_i(W_{\mathcal{S}})| = \frac{1}{2}\left((W_{L:1}(\bar{t}))_{1,1} + \sqrt{(W_{L:1}(\bar{t}))_{1,1}^2 + 4}\right)\,.$$

Thus, lower bounding $(W_{L:1}(\bar{t}))_{1,1}$ with Equation (75), and noticing that $(W_{L:1}(\bar{t}))_{1,1}^2 + 4 \geq 16$, leads to $\sigma_1(W_{\mathcal{S}}) \geq R^L/2^{L+2}$. Analogously, for the smallest singular value of $W_{\mathcal{S}}$ we have:

$$\begin{aligned} \sigma_2(W_{\mathcal{S}}) &= \min_{i=1,2} |\lambda_i(W_{\mathcal{S}})| \\ &= \frac{1}{2}\left(\sqrt{(W_{L:1}(\bar{t}))_{1,1}^2 + 4} - (W_{L:1}(\bar{t}))_{1,1}\right) \\ &= \frac{2}{\sqrt{(W_{L:1}(\bar{t}))_{1,1}^2 + 4} + (W_{L:1}(\bar{t}))_{1,1}} \\ &\leq 2^{L+2}/R^L\,, \end{aligned}$$

where in the third transition we made use of the identity $a - b = \frac{a^2-b^2}{a+b}$ for $a, b \in \mathbb{R}$ such that $a + b \neq 0$. Corollary 8.6.2 in [33] yields the following singular values perturbation bound:

$$|\sigma_i(W_{L:1}(\bar{t})) - \sigma_i(W_{\mathcal{S}})| \leq \left\| W_{L:1}(\bar{t}) - W_{\mathcal{S}} \right\|_F \quad , \; i = 1, 2\,.$$

Equation (80) then readily follows from the fact that $\left\| W_{L:1}(\bar{t}) - W_{\mathcal{S}} \right\|_F = \sqrt{2\ell(W_{L:1}(\bar{t}))}$.

By similar arguments, Equation (80) holds when $(W_{L:1}(\bar{t}))_{1,1} < 0$ as well. □