Implicit Regularization in Deep Matrix Factorization

Sanjeev Arora
Princeton University and Institute for Advanced Study
arora@cs.princeton.edu

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

Wei Hu
Princeton University
huwei@cs.princeton.edu

Yuping Luo
Princeton University
yupingl@cs.princeton.edu

Abstract

Efforts to understand the generalization mystery in deep learning have led to the belief that gradient-based optimization induces a form of implicit regularization, a bias towards models of low "complexity." We study the implicit regularization of gradient descent over deep linear neural networks for matrix completion and sensing, a model referred to as deep matrix factorization. Our first finding, supported by theory and experiments, is that adding depth to a matrix factorization enhances an implicit tendency towards low-rank solutions, oftentimes leading to more accurate recovery. Secondly, we present theoretical and empirical arguments questioning a nascent view by which implicit regularization in matrix factorization can be captured using simple mathematical norms. Our results point to the possibility that the language of standard regularizers may not be rich enough to fully encompass the implicit regularization brought forth by gradient-based optimization.

1   Introduction

It is a mystery how deep neural networks generalize despite having far more learnable parameters than training examples. Explicit regularization techniques alone cannot account for this generalization, as they do not prevent the networks from being able to fit random data (see [52]). A view by which gradient-based optimization induces an *implicit regularization* has thus arisen. Of course, this view would be uninsightful if "implicit regularization" were treated as synonymous with "promoting generalization" — the question is whether we can characterize the implicit regularization independently of any validation data. Notably, the mere use of the term "regularization" already predisposes us towards characterizations based on known explicit regularizers (*e.g.* a constraint on some norm of the parameters), but one must also be open to the possibility that something else is afoot.

An old argument (*cf.* [25, 29]) traces implicit regularization in deep learning to beneficial effects of noise introduced by small-batch stochastic optimization. The feeling is that solutions that do not generalize correspond to "sharp minima," and added noise prevents convergence to such solutions. However, recent evidence (*e.g.* [26, 51]) suggests that deterministic (or near-deterministic) gradient-based algorithms can also generalize, and thus a different explanation is in order.

A major hurdle in this study is that implicit regularization in deep learning seems to kick in only with certain types of data (not with random data for example), and we lack mathematical tools for reasoning about real-life data. Thus one needs a simple test-bed for the investigation, where data admits a crisp mathematical formulation. Following earlier works, we focus on the problem of *matrix completion*: given a randomly chosen subset of entries from an unknown matrix W^*, the task is to recover the unseen entries. To cast this as a prediction problem, we may view each entry in W^* as 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 [11] 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
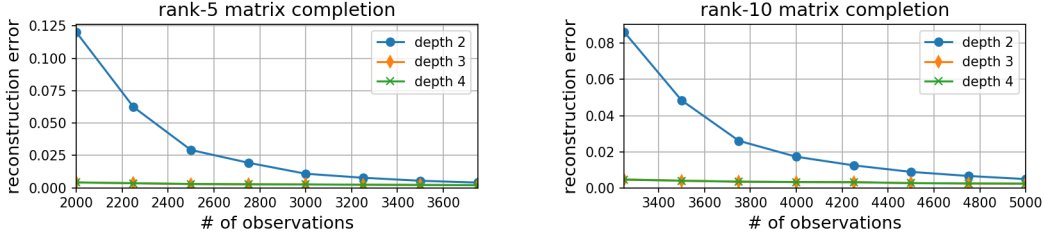
Figure 1: Matrix completion via gradient descent over deep matrix factorizations. Left (respectively, right) plot shows reconstruction errors for matrix factorizations of depths 2, 3 and 4, when applied to the completion of a random rank-5 (respectively, rank-10) matrix with size 100×100. x-axis stands for the number of observed entries (randomly chosen), y-axis represents reconstruction error, and error bars (indiscernible) mark standard deviations of the results over multiple trials. All matrix factorizations are full-dimensional, *i.e.* have hidden dimensions 100. Both learning rate and standard deviation of (random, zero-centered) initialization for gradient descent were set to the small value 10^{-3}. Notice, with few observed entries factorizations of depths 3 and 4 significantly outperform that of depth 2, whereas with more entries all factorizations perform well. For further details, and a similar experiment on matrix sensing tasks, see Appendix E.

can achieve approximate or even exact recovery. Of these, a well-known method based upon convex optimization finds the minimal nuclear norm matrix among those fitting all observed entries (see [9]).[1]

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 — $W = W_2 W_1$ — and optimizing the resulting (non-convex) objective for fitting observed entries. 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, in practice, even when the rank is unconstrained, running gradient descent with small learning rate (step size) and initialization close to zero tends to produce low-rank solutions, and thus allows accurate recovery if W^* is low-rank. This empirical observation led Gunasekar *et al.* to conjecture in [20] that gradient descent over a matrix factorization induces an implicit regularization minimizing nuclear norm:

Conjecture 1 (from [20], informally stated). *With small enough learning rate and initialization close enough to the origin, gradient descent on a full-dimensional matrix factorization converges to the minimum nuclear norm solution.*

Deep matrix factorization Since standard matrix factorization can be viewed as a two-layer neural network (with linear activations), a natural extension is to consider deeper models. A *deep matrix factorization*[2] of $W \in \mathbb{R}^{d,d'}$, with hidden dimensions $d_1, \ldots, d_{N-1} \in \mathbb{N}$, is the parameterization:

$$W = W_N W_{N-1} \cdots W_1 \,, \tag{1}$$

where $W_j \in \mathbb{R}^{d_j, d_{j-1}}$, $j = 1, \ldots, N$, with $d_N := d, d_0 := d'$. N is referred to as the *depth* of the factorization, the matrices $W_1, \ldots, W_N$ as its *factors*, and the resulting W as the *product matrix*.

Could the implicit regularization of deep matrix factorizations be stronger than that of their shallow counterpart (which Conjecture 1 equates with nuclear norm minimization)? Experiments reported in Figure 1 suggest that this is indeed the case — depth leads to more accurate completion of a low-rank matrix when the number of observed entries is small. Our purpose in the current paper is to mathematically analyze this stronger form of implicit regularization. Can it be described by a matrix norm (or quasi-norm) continuing the line of Conjecture 1, or is a paradigm shift required?

1.1 Paper overview

Review of related work is given in Appendix A (deferred to supplementary material per lack of space).

In Section 2 we investigate the potential of norms for capturing the implicit regularization in deep matrix factorization. Surprisingly, we find that the main theoretical evidence connecting nuclear norm and shallow (depth-2) matrix factorization — proof given in [20] for Conjecture 1 in a particular restricted setting — extends to arbitrarily deep factorizations as well. This result disqualifies the

---

[1]Recall that 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.

[2]Note that the literature includes various usages of this term — some in line with ours (*e.g.* [47, 53, 33]), while others less so (*e.g.* [50, 16, 49]).

natural hypothesis by which Schatten quasi-norms replace nuclear norm as the implicit regularization when one adds depth to a shallow matrix factorization. Instead, when interpreted through the lens of [20], it brings forth a conjecture by which the implicit regularization is captured by nuclear norm for any depth. Since our experiments (Figure 1) show that depth changes (enhances) the implicit regularization, we are led to question the theoretical direction proposed in [20], and accordingly conduct additional experiments to evaluate the validity of Conjecture 1.

Typically, when the number of observed entries is sufficiently large with respect to the rank of the matrix to recover, nuclear norm minimization yields exact recovery, and thus it is impossible to distinguish between that and a different implicit regularization which also perfectly recovers. The regime most interesting to evaluate is therefore that in which the number of observed entries is too small for exact recovery by nuclear norm minimization — here there is room for different implicit regularizations to manifest themselves by providing higher quality solutions. Our empirical results show that in this regime, matrix factorizations consistently outperform nuclear norm minimization, suggesting that their implicit regularization admits stronger bias towards low-rank, in contrast to Conjecture 1. Together, our theory and experiments lead us to suspect that the implicit regularization in matrix factorization (shallow or deep) may not be amenable to description by a simple mathematical norm, and a detailed analysis of the dynamics in optimization may be necessary.

Section 3 carries out such an analysis, characterizing how the singular value decomposition of the learned solution evolves during gradient descent. Evolution rates of singular values turn out to be proportional to their size exponentiated by $2 - 2/N$, where N is the depth of the factorization. This establishes a tendency towards low rank solutions, which intensifies with depth. Experiments validate the findings, demonstrating the dynamic nature of implicit regularization in deep matrix factorization.

We believe the trajectories traversed in optimization may be key to understanding generalization in deep learning, and hope that our work will inspire further progress along this line.

2 Can the implicit regularization be captured by norms?

In this section we investigate the possibility of extending Conjecture 1 for explaining implicit regularization in deep matrix factorization. Given the experimental evidence in Figure 1, one may hypothesize that gradient descent on a depth-N matrix factorization implicitly minimizes some norm (or quasi-norm) that approximates rank, with the approximation being more accurate the larger N is. For example, a natural candidate would be Schatten-p quasi-norm to the power of p ($0 < p \leq 1$), which for a matrix $W \in \mathbb{R}^{d,d'}$ is defined as: $\|W\|_{S_p}^p := \sum_{r=1}^{\min\{d,d'\}} \sigma_r^p(W)$, where $\sigma_1(W), \ldots, \sigma_{\min\{d,d'\}}(W)$ are the singular values of W. For $p = 1$ this reduces to nuclear norm, which by Conjecture 1 corresponds to a depth-2 factorization. As p approaches zero we obtain a closer approximation of $\text{rank}(W)$, which could be suitable for factorizations of higher depths.

We will focus in this section on *matrix sensing* — a more general problem than matrix completion. Here, we are given m measurement matrices $A_1, \ldots, A_m$, with corresponding labels $y_1, \ldots, y_m$ generated by $y_i = \langle A_i, W^* \rangle$, and our goal is to reconstruct the unknown matrix W^*. As in the case of matrix completion, well-known methods, and in particular nuclear norm minimization, can recover W^* if it is low-rank, certain technical conditions are met, and sufficiently many observations are given (see [42]).

2.1 Current theory does not distinguish depth-N from depth-2

Our first result is that the theory developed by [20] to support Conjecture 1 can be generalized to suggest that nuclear norm captures the implicit regularization in matrix factorization not just for depth 2, but for arbitrary depth. This is of course inconsistent with the experimental findings reported in Figure 1. We will first recall the existing theory, and then show how to extend it.

[20] studied implicit regularization in shallow (depth-2) matrix factorization by considering recovery of a positive semidefinite matrix from sensing via symmetric measurements, namely:

$$\min_{W \in \mathcal{S}_+^d} \ell(W) := \tfrac{1}{2} \sum_{i=1}^{m} (y_i - \langle A_i, W \rangle)^2, \tag{2}$$

where $A_1, \ldots, A_m \in \mathbb{R}^{d,d}$ are symmetric and linearly independent, and $\mathcal{S}_+^d$ stands for the set of (symmetric and) positive semidefinite matrices in $\mathbb{R}^{d,d}$. Focusing on the underdetermined regime $m \ll d^2$, they investigated the implicit bias brought forth by running *gradient flow* (gradient descent with infinitesimally small learning rate) on a symmetric full-rank matrix factorization, *i.e.* on the objective:

$$\psi : \mathbb{R}^{d,d} \to \mathbb{R}_{\geq 0} \quad , \quad \psi(Z) := \ell(ZZ^\top) = \tfrac{1}{2}\sum\nolimits_{i=1}^m (y_i - \langle A_i, ZZ^\top\rangle)^2 \, .$$

For $\alpha > 0$, denote by $W_{\mathrm{sha},\infty}(\alpha)$ (sha here stands for "shallow") the final solution $ZZ^\top$ obtained from running gradient flow on $\psi(\cdot)$ with initialization αI (α times identity). Formally, $W_{\mathrm{sha},\infty}(\alpha) := \lim_{t\to\infty} Z(t)Z(t)^\top$ where $Z(0) = \alpha I$ and $\dot{Z}(t) = -\frac{d\psi}{dZ}(Z(t))$ for $t \in \mathbb{R}_{\geq 0}$ (t here is a continuous time index, and $\dot{Z}(t)$ stands for the derivative of $Z(t)$ with respect to time). Letting $\|\cdot\|_*$ represent matrix nuclear norm, the following result was proven by [20]:

Theorem 1 (adaptation of Theorem 1 in [20]). *Assume the measurement matrices $A_1, \ldots, A_m$ commute. Then, if $\bar{W}_{\mathrm{sha}} := \lim_{\alpha\to 0} W_{\mathrm{sha},\infty}(\alpha)$ exists and is a global optimum for Equation (2) with $\ell(\bar{W}_{\mathrm{sha}}) = 0$, it holds that $\bar{W}_{\mathrm{sha}} \in \mathrm{argmin}_{W\in\mathcal{S}_+^d, \, \ell(W)=0} \|W\|_*$, i.e. $\bar{W}_{\mathrm{sha}}$ is a global optimum with minimal nuclear norm.*[3]

Motivated by Theorem 1 and empirical evidence they provided, [20] raised Conjecture 1, which, formally stated, hypothesizes that the condition in Theorem 1 of $\{A_i\}_{i=1}^m$ commuting is unnecessary, and an identical statement holds for arbitrary (symmetric linearly independent) measurement matrices.[4]

While the analysis of [20] covers only symmetric matrix factorizations of the form $ZZ^\top$, they noted that it can be extended to also account for asymmetric factorizations of the type considered in the current paper. Specifically, running gradient flow on the objective:

$$\phi(W_1, W_2) := \ell(W_2 W_1) = \tfrac{1}{2}\sum\nolimits_{i=1}^m (y_i - \langle A_i, W_2 W_1\rangle)^2 \, ,$$

with $W_1, W_2 \in \mathbb{R}^{d,d}$ initialized to αI, $\alpha > 0$, and denoting by $W_{\mathrm{sha},\infty}(\alpha)$ the product matrix obtained at the end of optimization (*i.e.* $W_{\mathrm{sha},\infty}(\alpha) := \lim_{t\to\infty} W_2(t)W_1(t)$ where $W_j(0) = \alpha I$ and $\dot{W}_j(t) = -\frac{\partial\phi}{\partial W_j}(W_1(t), W_2(t))$ for $t \in \mathbb{R}_{\geq 0}$), Theorem 1 holds exactly as stated. For completeness, we provide a proof of this fact in Appendix D.

Next, we show that Theorem 1 — the main theoretical justification for the connection between nuclear norm and shallow matrix factorization — extends to arbitrarily deep factorizations as well. Consider gradient flow over the objective:

$$\phi(W_1, \ldots, W_N) := \ell(W_N W_{N-1}\cdots W_1) = \tfrac{1}{2}\sum\nolimits_{i=1}^m (y_i - \langle A_i, W_N W_{N-1}\cdots W_1\rangle)^2 \, ,$$

with $W_1, \ldots, W_N \in \mathbb{R}^{d,d}$ initialized to αI, $\alpha > 0$. Using $W_{\mathrm{deep},\infty}(\alpha)$ to denote the product matrix obtained at the end of optimization (*i.e.* $W_{\mathrm{deep},\infty}(\alpha) := \lim_{t\to\infty} W_N(t)W_{N-1}(t)\cdots W_1(t)$ where $W_j(0) = \alpha I$ and $\dot{W}_j(t) = -\frac{\partial\phi}{\partial W_j}(W_1(t), \ldots, W_N(t))$ for $t \in \mathbb{R}_{\geq 0}$), a result analogous to Theorem 1 holds:

Theorem 2. *Suppose $N \geq 3$, and that the matrices $A_1, \ldots, A_m$ commute. Then, if $\bar{W}_{\mathrm{deep}} := \lim_{\alpha\to 0} W_{\mathrm{deep},\infty}(\alpha)$ exists and is a global optimum for Equation (2) with $\ell(\bar{W}_{\mathrm{deep}}) = 0$, it holds that $\bar{W}_{\mathrm{deep}} \in \mathrm{argmin}_{W\in\mathcal{S}_+^d, \, \ell(W)=0} \|W\|_*$, i.e. $\bar{W}_{\mathrm{deep}}$ is a global optimum with minimal nuclear norm.*[5]

Proof sketch (for complete proof see Appendix C.1). Our proof is inspired by that of Theorem 1 given in [20]. Using the expression for $\dot{W}(t)$ derived in [3] (Lemma 3 in Appendix B), it can be shown that $W(t)$ commutes with $\{A_i\}_{i=1}^m$, and takes on a particular form. Taking limits $t \to \infty$ and $\alpha \to 0$, optimality (minimality) of nuclear norm is then established using a duality argument. $\quad\square$

Theorem 2 provides a particular setting where the implicit regularization in deep matrix factorizations boils down to nuclear norm minimization. By Proposition 1 below, there exist instances of this setting for which the minimization of nuclear norm contradicts minimization (even locally) of Schatten-p

---

[3]The result of [20] is slightly more general — it allows gradient flow to be initialized by αO, where O is an arbitrary orthogonal matrix, and it is shown that this leads to the exact same $W_{\mathrm{sha},\infty}(\alpha)$ as one would obtain from initializing at αI. For simplicity, we limit our discussion to the latter initialization.

[4]Their conjecture also relaxes the requirement from the initialization of gradient flow — an initial value of αZ_0 is believed to suffice, where Z_0 is an arbitrary full-rank matrix (that does not depend on α).

[5]By Appendix C.1: $W_N(t)W_{N-1}(t)\cdots W_1(t) \succeq 0 \;\forall t$. Therefore, even though the theorem treats optimization over $\mathcal{S}_+^d$ using an unconstrained asymmetric factorization, gradient flow implicitly constrains the search to $\mathcal{S}_+^d$, so the assumption of $\bar{W}_{\mathrm{deep}}$ being a global optimum for Equation (2) with $\ell(\bar{W}_{\mathrm{deep}}) = 0$ is no stronger than the analogous assumption in Theorem 1 from [20]. The implicit constraining to $\mathcal{S}_+^d$ also holds when $N = 2$ (see Appendix D), so the asymmetric extension of Theorem 1 does not involve strengthening assumptions either.
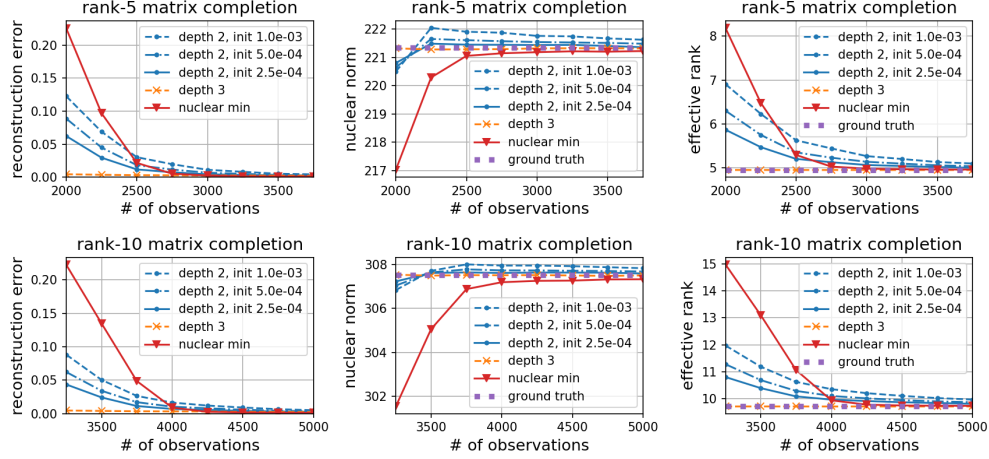
Figure 2: Evaluation of nuclear norm as the implicit regularization in deep matrix factorization. Each plot compares gradient descent over matrix factorizations of depths 2 and 3 (results for depth 4 were indistinguishable from those of depth 3; we omit them to reduce clutter) against minimum nuclear norm solution and ground truth in matrix completion tasks. Top (respectively, bottom) row corresponds to completion of a random rank-5 (respectively, rank-10) matrix with size 100×100. Left, middle and right columns display (in y-axis) reconstruction error, nuclear norm and effective rank (*cf.* [43]) respectively. In each plot, x-axis stands for the number of observed entries (randomly chosen), and error bars (indiscernible) mark standard deviations of the results over multiple trials. All matrix factorizations are full-dimensional, *i.e.* have hidden dimensions 100. Both learning rate and standard deviation of (random, zero-centered) initialization for gradient descent were initially set to 10^{-3}. Running with smaller learning rate did not yield a noticeable change in terms of final results. Initializing with smaller standard deviation had no observable effect on results of depth 3 (and 4), but did impact those of depth 2 — the outcomes of dividing standard deviation by 2 and by 4 are included in the plots.[7] Notice, with many observed entries minimum nuclear norm solution coincides with ground truth (minimum rank solution), and matrix factorizations of all depths converge to these. On the other hand, when there are fewer observed entries minimum nuclear norm solution does not coincide with ground truth, and matrix factorizations prefer to lower the effective rank at the expense of higher nuclear norm, in a manner that is more potent for deeper factorizations. For further details, and a similar experiment on matrix sensing tasks, see Appendix E.

quasi-norm for any $0 < p < 1$. Therefore, one cannot hope to capture the implicit regularization in deep matrix factorizations through Schatten quasi-norms. Instead, if we interpret Theorem 2 through the lens of [20], we arrive at a conjecture by which the implicit regularization is captured by nuclear norm for any depth.

Proposition 1. *For any dimension $d \geq 3$, there exist linearly independent symmetric and commutable measurement matrices $A_1, \ldots, A_m \in \mathbb{R}^{d,d}$, and corresponding labels $y_1, \ldots, y_m \in \mathbb{R}$, such that the limit solution defined in Theorem 2 — $\bar{W}_{\text{deep}}$ — which has been shown to satisfy $\bar{W}_{\text{deep}} \in \arg\min_{W \in \mathcal{S}_+^d, \ell(W)=0} \|W\|_*$, is not a local minimum of the following program for any $0 < p < 1$:[6]*
$$\min_{W \in \mathcal{S}_+^d, \ell(W)=0} \|W\|_{S_p}.$$
Proof sketch (for complete proof see Appendix C.2). We choose $A_1, \ldots, A_m$ and $y_1, \ldots, y_m$ such that: *(i)* $\bar{W}_{\text{deep}} = \text{diag}(1, 1, 0, \ldots, 0)$; and *(ii)* adding $\epsilon \in (0, 1)$ to entries $(1, 2)$ and $(2, 1)$ of $\bar{W}_{\text{deep}}$ maintains optimality. The result then follows from the fact that the addition of ϵ decreases Schatten-p quasi-norm for any $0 < p < 1$. $\square$

2.2 Experiments challenging Conjecture 1

Subsection 2.1 suggests that from the perspective of current theory, it is natural to apply Conjecture 1 to matrix factorizations of arbitrary depth. On the other hand, the experiment reported in Figure 1 implies that depth changes (enhances) the implicit regularization. To resolve this tension we conduct a more refined experiment, which ultimately puts in question the validity of Conjecture 1.

---

[6]Following [20], we take for granted existence of $\bar{W}_{\text{deep}}$ and it being a global optimum for Equation (2) with $\ell(\bar{W}_{\text{deep}}) = 0$. If this is not the case then Theorem 2 does not apply, and hence it obviously does not disqualify minimization of Schatten quasi-norms as the implicit regularization.

[7]As can be seen, using smaller initialization enhanced the implicit tendency of depth-2 matrix factorization towards low rank. It is possible that this tendency can eventually match that of depth-3 (and -4), but only if initialization size goes far below what is customary in deep learning.

Our experimental protocol is as follows. For different matrix completion tasks with varying number of observed entries, we compare minimum nuclear norm solution to those brought forth by running gradient descent on matrix factorizations of different depths. For each depth, we apply gradient descent with different choices of learning rate and standard deviation for (random, zero-centered) initialization, observing the trends as these become smaller. The outcome of the experiment is presented in Figure 2. As can be seen, when the number of observed entries is sufficiently large with respect to the rank of the matrix to recover, factorizations of all depths indeed admit solutions that tend to minimum nuclear norm. However, when there are less entries observed — precisely the data-poor setting where implicit regularization matters most — neither shallow (depth-2) nor deep (depth-N with $N \geq 3$) factorizations minimize nuclear norm. Instead, they put more emphasis on lowering the effective rank (*cf.* [43]), in a manner which is stronger for deeper factorizations.

A close look at the experiments of [20] reveals that there too, in situations where the number of observed entries (or sensing measurements) was small (less than required for reliable recovery), a discernible gap appeared between the minimal nuclear norm and that returned by (gradient descent on) a matrix factorization. In light of Figure 2, we believe that if [20] had included in its plots an accurate surrogate for rank (*e.g.* effective rank or Schatten-p quasi-norm with small p), scenarios where matrix factorization produced sub-optimal (higher than minimum) nuclear norm would have manifested superior (lower) rank. More broadly, our experiments suggest that the implicit regularization in (shallow or deep) matrix factorization is somehow geared towards low rank, and just so happens to minimize nuclear norm in cases with sufficiently many observations, where minimum nuclear norm and minimum rank are known to coincide (*cf.* [9, 42]). We note that the theoretical analysis of [32] supporting Conjecture 1 is limited to such cases, and thus cannot truly distinguish between nuclear norm minimization and some other form of implicit regularization favoring low rank.

Given that Conjecture 1 seems to hold in some settings (Theorems 1 and 2; [32]) but not in other (Figure 2), we hypothesize that capturing implicit regularization in (shallow or deep) matrix factorization through a single mathematical norm (or quasi-norm) may not be possible, and a detailed account for the optimization process might be necessary. This is carried out in Section 3.

3 Dynamical analysis

This section characterizes trajectories of gradient flow (gradient descent with infinitesimally small learning rate) on deep matrix factorizations. The characterization significantly extends past analyses for linear neural networks (*e.g.* [44, 3]) — we derive differential equations governing the dynamics of singular values and singular vectors for the product matrix W (Equation (1)). Evolution rates of singular values turn out to be proportional to their size exponentiated by $2 - 2/N$, where N is the depth of the factorization. For singular vectors, we show that lack of movement implies a particular form of alignment with the gradient, and by this strengthen past results which have only established the converse. Via theoretical and empirical demonstrations, we explain how our findings imply a tendency towards low-rank solutions, which intensifies with depth.

Our derivation treats a setting which includes matrix completion and sensing as special cases. We assume minimization of a general analytic loss $\ell(\cdot)$,[8] overparameterized by a deep matrix factorization:

$$\phi(W_1, \ldots, W_N) := \ell(W_N W_{N-1} \cdots W_1). \tag{3}$$

We study gradient flow over the factorization:

$$\dot{W}_j(t) := \tfrac{d}{dt} W_j(t) = -\tfrac{\partial}{\partial W_j} \phi(W_1(t), \ldots, W_N(t)) \quad , \ t \geq 0 \, , \ j = 1, \ldots, N \, , \tag{4}$$

and in accordance with past work, assume that factors are *balanced* at initialization, *i.e.*:

$$W_{j+1}^\top(0) W_{j+1}(0) = W_j(0) W_j^\top(0) \quad , \ j = 1, \ldots, N - 1 \, . \tag{5}$$

Equation (5) is satisfied approximately in the common setting of near-zero initialization (it holds exactly in the "residual" setting of identity initialization — *cf.* [23, 5]). The condition played an important role in the analysis of [3], facilitating derivation of a differential equation governing the product matrix of a linear neural network (see Lemma 3 in Appendix B). It was shown in [3] empirically 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.* [4, 28]) later supported this match theoretically.

---

[8]A function $f(\cdot)$ is *analytic* on a domain $\mathcal{D}$ if at every $x \in \mathcal{D}$: it is infinitely differentiable; and its Taylor series converges to it on some neighborhood of x (see [30] for further details).

We note that by Section 6 in [3], for depth $N \geq 3$, the dynamics of the product matrix W (Equation (1)) *cannot* be exactly equivalent to gradient descent on the loss $\ell(\cdot)$ regularized by a penalty term. This preliminary observation already hints to the possibility that the effect of depth is different from those of standard regularization techniques.

Employing results of [3], we will characterize the evolution of singular values and singular vectors for W. As a first step, we show that W admits an *analytic singular value decomposition* ([7, 12]):

Lemma 1. *The product matrix $W(t)$ can be expressed as:*

$$W(t) = U(t)S(t)V^\top(t), \tag{6}$$

where: $U(t) \in \mathbb{R}^{d,\min\{d,d'\}}$, $S(t) \in \mathbb{R}^{\min\{d,d'\},\min\{d,d'\}}$ and $V(t) \in \mathbb{R}^{d',\min\{d,d'\}}$ are analytic functions of t; and for every t, the matrices $U(t)$ and $V(t)$ have orthonormal columns, while $S(t)$ is diagonal (elements on its diagonal may be negative and may appear in any order).

Proof sketch (for complete proof see Appendix C.3). We show that $W(t)$ is an analytic function of t and then invoke Theorem 1 in [7]. □

The diagonal elements of $S(t)$, which we denote by $\sigma_1(t), \ldots, \sigma_{\min\{d,d'\}}(t)$, are signed singular values of $W(t)$; the columns of $U(t)$ and $V(t)$, denoted $\mathbf{u}_1(t), \ldots, \mathbf{u}_{\min\{d,d'\}}(t)$ and $\mathbf{v}_1(t), \ldots, \mathbf{v}_{\min\{d,d'\}}(t)$, are the corresponding left and right singular vectors (respectively).

With Lemma 1 in place, we are ready to characterize the evolution of singular values:

Theorem 3. *The signed singular values of the product matrix $W(t)$ evolve by:*

$$\dot{\sigma}_r(t) = -N \cdot \left(\sigma_r^2(t)\right)^{1-1/N} \cdot \left\langle \nabla\ell(W(t)), \mathbf{u}_r(t)\mathbf{v}_r^\top(t) \right\rangle \quad, \ r = 1, \ldots, \min\{d, d'\}. \tag{7}$$

If the matrix factorization is non-degenerate, i.e. has depth $N \geq 2$, the singular values need not be signed (we may assume $\sigma_r(t) \geq 0$ for all t).

Proof sketch (for complete proof see Appendix C.4). Differentiating the analytic singular value decomposition (Equation (6)) with respect to time, multiplying from the left by $U^\top(t)$ and from the right by $V(t)$, and using the fact that $U(t)$ and $V(t)$ have orthonormal columns, we obtain $\dot{\sigma}_r(t) = \mathbf{u}_r^\top(t)\dot{W}(t)\mathbf{v}_r(t)$. Equation (7) then follows from plugging in the expression for $\dot{W}(t)$ developed by [3] (Lemma 3 in Appendix B). □

Strikingly, given a value for $W(t)$, the evolution of singular values depends on N — depth of the matrix factorization — only through the multiplicative factors $N \cdot (\sigma_r^2(t))^{1-1/N}$ (see Equation (7)). In the degenerate case $N = 1$, *i.e.* when the product matrix $W(t)$ is simply driven by gradient flow over the loss $\ell(\cdot)$ (no matrix factorization), the multiplicative factors reduce to 1, and the singular values evolve by: $\dot{\sigma}_r(t) = -\left\langle \nabla\ell(W(t)), \mathbf{u}_r(t)\mathbf{v}_r^\top(t) \right\rangle$. With $N \geq 2$, *i.e.* when depth is added to the factorization, the multiplicative factors become non-trivial, and while the constant N does not differentiate between singular values, the terms $(\sigma_r^2(t))^{1-1/N}$ do — they enhance movement of large singular values, and on the other hand attenuate that of small ones. Moreover, the enhancement/attenuation becomes more significant as N (depth of the factorization) grows.

Next, we turn to the evolution of singular vectors:

Lemma 2. *Assume that at initialization, the singular values of the product matrix $W(t)$ are distinct and different from zero.[9] Then, its singular vectors evolve by:*

$$\dot{U}(t) = -U(t)\left(F(t) \odot \left[U^\top(t)\nabla\ell(W(t))V(t)S(t) + S(t)V^\top(t)\nabla\ell^\top(W(t))U(t)\right]\right)$$
$$- \left(I_d - U(t)U^\top(t)\right)\nabla\ell(W(t))V(t)(S^2(t))^{\frac{1}{2}-\frac{1}{N}} \tag{8}$$
$$\dot{V}(t) = -V(t)\left(F(t) \odot \left[S(t)U^\top(t)\nabla\ell(W(t))V(t) + V^\top(t)\nabla\ell^\top(W(t))U(t)S(t)\right]\right)$$
$$- \left(I_{d'} - V(t)V^\top(t)\right)\nabla\ell^\top(W(t))U^\top(t)(S^2(t))^{\frac{1}{2}-\frac{1}{N}}, \tag{9}$$

where I_d and $I_{d'}$ are the identity matrices of sizes $d \times d$ and $d' \times d'$ respectively, $\odot$ stands for the Hadamard (element-wise) product, and the matrix $F(t) \in \mathbb{R}^{\min\{d,d'\},\min\{d,d'\}}$ is skew-symmetric with $((\sigma_{r'}^2(t))^{1/N} - (\sigma_r^2(t))^{1/N})^{-1}$ in its (r, r')'th entry, $r \neq r'$.[10]

---

[9] This assumption can be relaxed significantly — all that is needed is that no singular value be identically zero ($\forall r \ \exists t \ s.t. \ \sigma_r(t) \neq 0$), and no pair of singular values be identical through time ($\forall r, r' \ \exists t \ s.t. \ \sigma_r(t) \neq \sigma_{r'}(t)$).

[10] Equations (8) and (9) are well-defined when t is such that $\sigma_1(t), \ldots, \sigma_{\min\{d,d'\}}(t)$ are distinct and different from zero. By analyticity, this is either the case for every t besides a set of isolated points, or it is not the case for any t. Our assumption on initialization disqualifies the latter option, so any t for which Equations (8) or (9) are ill-defined is isolated. The derivatives of U and V for such t may thus be inferred by continuity.
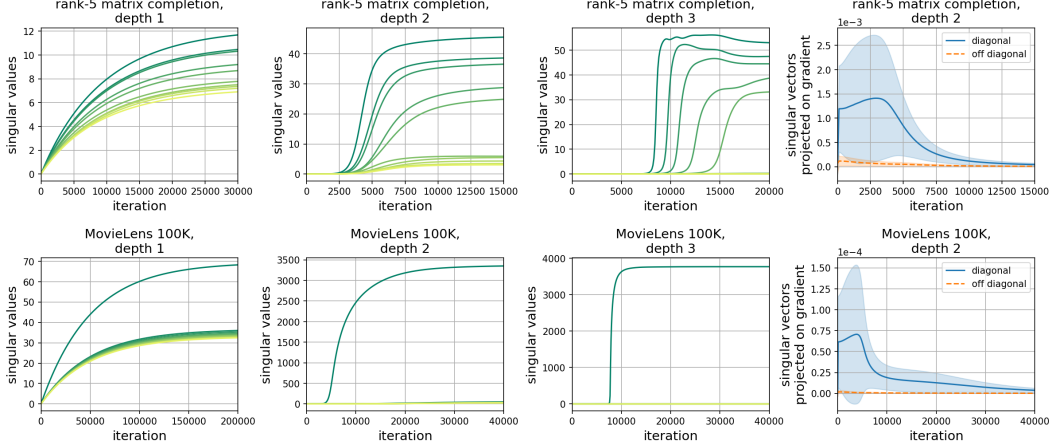
Figure 3: Dynamics of gradient descent over deep matrix factorizations — specifically, evolution of singular values and singular vectors of the product matrix during training for matrix completion. Top row corresponds to the task of completing a random rank-5 matrix with size 100×100 based on 2000 randomly chosen observed entries; bottom row corresponds to training on 10000 entries chosen randomly from the MovieLens 100K dataset (completion of a 943×1682 matrix, *cf.* [24]).[11] First (left) three columns show top singular values for, respectively, depths 1 (no matrix factorization), 2 (shallow matrix factorization) and 3 (deep matrix factorization). Last (right) column shows singular vectors for a depth-2 factorization, by comparing on- *vs.* off-diagonal entries in the matrix $U^\top(t)\nabla\ell(W(t))V(t)$ (see Corollary 1) — for each group of entries, mean of absolute values is plotted, along with shaded area marking the standard deviation. All matrix factorizations are full-dimensional (hidden dimensions 100 in top row plots, 943 in bottom row plots). Notice, increasing depth makes singular values move slower when small and faster when large (in accordance with Theorem 3), which results in solutions with effectively lower rank. Notice also that $U^\top(t)\nabla\ell(W(t))V(t)$ is diagonally dominant so long as there is movement, showing that singular vectors of the product matrix align with those of the gradient (in accordance with Corollary 1). For further details, and a similar experiment on matrix sensing, see Appendix E.

Proof sketch (for complete proof see Appendix C.5). We follow a series of steps adopted from [46] to obtain expressions for $\dot{U}(t)$ and $\dot{V}(t)$ in terms of $U(t)$, $V(t)$, $S(t)$ and $\dot{W}(t)$. Plugging in the expression for $\dot{W}(t)$ developed by [3] (Lemma 3 in Appendix B) then yields Equations (8), (9). □

Corollary 1. *Assume the conditions of Lemma 2, and that the matrix factorization is non-degenerate, i.e. has depth $N \geq 2$. Then, for any time t such that the singular vectors of the product matrix $W(t)$ are stationary, i.e. $\dot{U}(t) = 0$ and $\dot{V}(t) = 0$, it holds that $U^\top(t)\nabla\ell(W(t))V(t)$ is diagonal, meaning they align with the singular vectors of $\nabla\ell(W(t))$.*

Proof sketch (for complete proof see Appendix C.6). By Equations (8) and (9), $U^\top(t)\dot{U}(t)S(t) - S(t)V^\top(t)\dot{V}(t)$ is equal to the Hadamard product between $U^\top(t)\nabla\ell(W(t))V(t)$ and a (time-dependent) square matrix with zeros on its diagonal and non-zeros elsewhere. When $\dot{U}(t) = 0$ and $\dot{V}(t) = 0$ obviously $U^\top(t)\dot{U}(t)S(t) - S(t)V^\top(t)\dot{V}(t) = 0$, and so the Hadamard product is zero. This implies that $U^\top(t)\nabla\ell(W(t))V(t)$ is diagonal. □

Earlier papers studying gradient flow for linear neural networks (*e.g.* [44, 1, 31]) could show that singular vectors are stationary if they align with the singular vectors of the gradient. Corollary 1 is significantly stronger and implies a converse — if singular vectors are stationary, they must be aligned with the gradient. Qualitatively, this suggests that a "goal" of gradient flow on a deep matrix factorization is to align singular vectors of the product matrix with those of the gradient.

3.1 Implicit regularization towards low rank

Figure 3 presents empirical demonstrations of our conclusions from Theorem 3 and Corollary 1. It shows that for a non-degenerate deep matrix factorization, *i.e.* one with depth $N \geq 2$, under gradient descent with small learning rate and near-zero initialization, singular values of the product matrix are subject to an enhancement/attenuation effect as described above: they progress very slowly after initialization, when close to zero; then, upon reaching a certain threshold, the movement of a singular

---

[11]Observations of MovieLens 100K were subsampled solely for reducing run-time.

value becomes rapid, with the transition from slow to rapid movement being sharper with a deeper factorization (larger N). In terms of singular vectors, the figure shows that those of the product matrix indeed align with those of the gradient. Overall, the dynamics promote solutions that have a few large singular values and many small ones, with a gap that is more extreme the deeper the matrix factorization is. This is an implicit regularization towards low rank, which intensifies with depth.

Theoretical illustration Consider the simple case of square matrix sensing with a single measurement fit via ℓ_2 loss: $\ell(W) = \frac{1}{2}(\langle A, W \rangle - y)^2$, where $A \in \mathbb{R}^{d,d}$ is the measurement matrix, and $y \in \mathbb{R}$ the corresponding label. Suppose we learn by running gradient flow over a depth-N matrix factorization, *i.e.* over the objective $\phi(\cdot)$ defined in Equation (3). Corollary 1 states that the singular vectors of the product matrix — $\{\mathbf{u}_r(t)\}_r$ and $\{\mathbf{v}_r(t)\}_r$ — are stationary only when they diagonalize the gradient, meaning $\{|\mathbf{u}_r^\top(t)\nabla\ell(W(t))\mathbf{v}_r| : r = 1, \ldots, d\}$ coincides with the set of singular values in $\nabla\ell(W(t))$. In our case $\nabla\ell(W) = (\langle A, W \rangle - y)A$, so stationarity of singular vectors implies $|\mathbf{u}_r^\top(t)\nabla\ell(W(t))\mathbf{v}_r| = |\delta(t)| \cdot \rho_r$, where $\delta(t) := \langle A, W(t) \rangle - y$ and $\rho_1, \ldots, \rho_d$ are the singular values of A (in no particular order). We will assume that starting from some time t_0 singular vectors are stationary, and accordingly $\mathbf{u}_r^\top(t)\nabla\ell(W(t))\mathbf{v}_r(t) = \delta(t) \cdot e_r \cdot \rho_r$ for $r = 1, \ldots, d$, where $e_1, \ldots, e_d \in \{-1, 1\}$. Theorem 3 then implies that (signed) singular values of the product matrix evolve by:

$$\dot{\sigma}_r(t) = -N \cdot \left(\sigma_r^2(t)\right)^{1-1/N} \cdot \delta(t) \cdot e_r \cdot \rho_r \quad , \forall t \geq t_0 \,. \tag{10}$$

Let $r_1, r_2 \in \{1, \ldots, d\}$. By Equation (10):

$$\int_{t'=t_0}^t \left(\sigma_{r_1}^2(t')\right)^{-1+1/N} \dot{\sigma}_{r_1}(t')dt' = \frac{e_{r_1}\rho_{r_1}}{e_{r_2}\rho_{r_2}} \cdot \int_{t'=t_0}^t \left(\sigma_{r_2}^2(t')\right)^{-1+1/N} \dot{\sigma}_{r_2}(t')dt' \,.$$

Computing the integrals, we may express $\sigma_{r_1}(t)$ as a function of $\sigma_{r_2}(t)$:[12]

$$\sigma_{r_1}(t) = \begin{cases} \alpha_{r_1,r_2} \cdot \sigma_{r_2}(t) + const & , N = 1 \\ \left(\sigma_{r_2}(t)\right)^{\alpha_{r_1,r_2}} \cdot const & , N = 2 \\ \left(\alpha_{r_1,r_2} \cdot \left(\sigma_{r_2}(t)\right)^{-\frac{N-2}{N}} + const\right)^{-\frac{N}{N-2}} & , N \geq 3 \end{cases} , \tag{11}$$

where $\alpha_{r_1,r_2} := e_{r_1}\rho_{r_1}(e_{r_2}\rho_{r_2})^{-1}$, and $const$ stands for a value that does not depend on t. Equation 11 reveals a gap between $\sigma_{r_1}(t)$ and $\sigma_{r_2}(t)$ that enhances with depth. For example, consider the case where $0 < \alpha_{r_1,r_2} < 1$. If the depth N is one, *i.e.* the matrix factorization is degenerate, $\sigma_{r_1}(t)$ will grow linearly with $\sigma_{r_2}(t)$. If $N = 2$ — shallow matrix factorization — $\sigma_{r_1}(t)$ will grow polynomially more slowly than $\sigma_{r_2}(t)$ (*const* here is positive). Increasing depth further will lead $\sigma_{r_1}(t)$ to asymptote when $\sigma_{r_2}(t)$ grows, at a value which can be shown to be lower the larger N is. Overall, adding depth to the matrix factorization leads to more significant gaps between singular values of the product matrix, *i.e.* to a stronger implicit bias towards low rank.

4   Conclusion

The implicit regularization of gradient-based optimization is key to generalization in deep learning. As a stepping stone towards understanding this phenomenon, we studied deep linear neural networks for matrix completion and sensing, a model referred to as deep matrix factorization. Through theory and experiments, we questioned prevalent norm-based explanations for implicit regularization in matrix factorization (*cf.* [20]), and offered an alternative, dynamical approach. Our characterization of the dynamics induced by gradient flow on the singular value decomposition of the learned matrix significantly extends prior work on linear neural networks. It reveals an implicit tendency towards low rank which intensifies with depth, supporting the empirical superiority of deeper matrix factorizations.

An emerging view is that understanding optimization in deep learning necessitates a detailed account for the trajectories traversed in training (*cf.* [4]). Our work adds another dimension to the potential importance of trajectories — we believe they are necessary for understanding generalization as well, and in particular, may be key to analyzing implicit regularization for non-linear neural networks.

---

[12]In accordance with Theorem 3, if $N \geq 2$, we assume without loss of generality that $\sigma_{r_1}(t), \sigma_{r_2}(t) \geq 0$, while disregarding the trivial case of equality to zero.

References

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

[2] Akshay Agrawal, Robin Verschueren, Steven Diamond, and Stephen Boyd. A rewriting system for convex optimization problems. *Journal of Control and Decision*, 5(1):42–60, 2018.

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

[4] 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*, 2019.

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

[6] Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, pages 3873–3881, 2016.

[7] Angelika Bunse-Gerstner, Ralph Byers, Volker Mehrmann, and Nancy K Nichols. Numerical computation of an analytic singular value decomposition of a matrix valued function. *Numerische Mathematik*, 60(1): 1–39, 1991.

[8] 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.

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

[10] Yuejie Chi, Yue M Lu, and Yuxin Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *arXiv preprint arXiv:1809.09573*, 2018.

[11] 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.

[12] B De Moor and S Boyd. Analytic properties of singular values and vectors. *Katholic Univ. Leuven, Belgium Tech. Rep*, 28:1989, 1989.

[13] Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

[14] Simon S Du and Wei Hu. Width provably matters in optimization for deep linear neural networks. *arXiv preprint arXiv:1901.08572*, 2019.

[15] Simon S Du, Wei Hu, and Jason D Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. *arXiv preprint arXiv:1806.00900*, 2018.

[16] Jicong Fan and Jieyu Cheng. Matrix completion by deep matrix factorization. *Neural Networks*, 98:34–41, 2018.

[17] Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981, 2016.

[18] Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1233–1242. JMLR. org, 2017.

[19] Gauthier Gidel, Francis Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in deep linear neural networks. *arXiv preprint arXiv:1904.13262*, 2019.

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

[21] 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*, volume 80, pages 1832–1841, 2018.

[22] 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*, pages 9461–9471, 2018.

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

[24] F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):19, 2016.

[25] Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Computation*, 9(1):1–42, 1997.

[26] Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. In *Advances in Neural Information Processing Systems*, pages 1731–1741, 2017.

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

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

[29] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *International Conference on Learning Representations*, 2017.

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

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

[32] 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*, pages 2–47, 2018.

[33] Zechao Li and Jinhui Tang. Deep matrix factorization for social image tag refinement and assignment. In *2015 IEEE 17th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6. IEEE, 2015.

[34] Junhong Lin, Raffaello Camoriano, and Lorenzo Rosasco. Generalization properties and implicit regularization for multiple passes sgm. In *International Conference on Machine Learning*, pages 2340–2348, 2016.

[35] 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*, pages 3351–3360, 2018.

[36] 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.

[37] 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.

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

[39] Dohyung Park, Anastasios Kyrillidis, Constantine Carmanis, and Sujay Sanghavi. Non-square matrix sensing without spurious local minima via the burer-monteiro approach. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 65–74, 2017.

[40] 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.

[41] 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. *arXiv preprint arXiv:1806.08734*, 2018.

[42] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[43] 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.

[44] 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*, 2014.

[45] 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.

[46] James Townsend. Differentiating the singular value decomposition. Technical report, 2016.

[47] George Trigeorgis, Konstantinos Bousmalis, Stefanos Zafeiriou, and Björn W Schuller. A deep matrix factorization method for learning attribute representations. *IEEE transactions on pattern analysis and machine intelligence*, 39(3):417–429, 2017.

[48] 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*, pages 964–973, 2016.

[49] Qi Wang, Mengying Sun, Liang Zhan, Paul Thompson, Shuiwang Ji, and Jiayu Zhou. Multi-modality disease modeling via collective deep matrix factorization. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1155–1164. ACM, 2017.

[50] Hong-Jian Xue, Xinyu Dai, Jianbing Zhang, Shujian Huang, and Jiajun Chen. Deep matrix factorization models for recommender systems. In *IJCAI*, pages 3203–3209, 2017.

[51] Yang You, Igor Gitman, and Boris Ginsburg. Scaling sgd batch size to 32k for imagenet training. *arXiv preprint arXiv:1708.03888*, 2017.

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

[53] Handong Zhao, Zhengming Ding, and Yun Fu. Multi-view clustering via deep matrix factorization. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

A Related work

Implicit regularization in deep learning is a highly active area of research. For non-linear neural networks, the topic has thus far been studied empirically (*e.g.* in [37, 52, 29, 26, 38]), with theoretical analyses being somewhat scarce (see [15, 41] for some of the few observations that have been derived). The majority of theoretical attention has been devoted to (single-layer) linear predictors and (multi-layer) linear neural networks, often viewed as stepping stones towards non-linear models. Linear predictors were treated in [34, 45, 36, 21]. For linear neural networks, [1, 31, 19] studied settings where the training objective admits a single global minimum, and the question is what path gradient descent (or gradient flow) takes to reach it.[13] This stands in contrast to the practical deep learning scenario where there are multiple global minima, and implicit regularization refers to the optimizer being biased towards reaching those solutions that generalize well. The latter scenario was treated by [22] and [28] in the context of linear neural networks trained for binary classification via separable data. These works showed that under certain assumptions, gradient descent converges (in direction) to the maximum margin solution. Intriguingly, the bias towards maximum margin holds with any number of layers, so in particular, implicit regularization was found to be oblivious to depth.[14]

The most extensively studied instance of linear neural networks is matrix factorization, corresponding to a model with multiple inputs, multiple outputs and a single hidden layer, typically trained to recover a low-rank linear mapping. The literature on matrix factorization for low-rank matrix recovery is far too broad to cover here — we refer to [10] for a recent survey, while mentioning that the technique is oftentimes attributed to [8]. Notable works proving successful recovery of a low-rank matrix through matrix factorization trained by gradient descent with no explicit regularization are [48, 35, 32]. Of these, [32] can be viewed as resolving the conjecture of [20] — which we investigate in Section 2 — for the case of sufficiently many linear measurements satisfying the restricted isometry property.

To the best of our knowledge, the current paper is the first to study implicit regularization for deep (three or more layer) linear neural networks with multiple outputs. The latter trait seems to be distinctive, as it is the main differentiator between the setting of [22, 28], where implicit regularization is oblivious to depth, and ours, for which we show that depth has significant impact. We note that our work is focused on the type of solutions reached by gradient descent, not the complementary questions of whether an optimal solution is found, and how fast that happens. These questions were studied extensively for matrix factorization — *cf.* [17, 6, 39, 18] — and more recently for linear neural networks of arbitrary depth — see [5, 3, 4, 14]. From a technical perspective, closest to our work are [20] and [3] — we rely on their results and significantly extend them (see Sections 2 and 3).

B Useful lemmas

We recall the following result from [3], which characterizes the evolution of the product matrix under gradient flow on a deep matrix factorization:

Lemma 3 (adaptation of Theorem 1 in [3]). *Let $\ell : \mathbb{R}^{d,d'} \to \mathbb{R}_{\geq 0}$ be a continuously differentiable loss, overparameterized by a deep matrix factorization:*

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

Suppose we run gradient flow over the factorization:

$$\dot{W}_j(t) := \tfrac{d}{dt} W_j(t) = -\tfrac{\partial}{\partial W_j} \phi(W_1(t), \ldots, W_N(t)) \quad , \, t \geq 0 \, , \, j = 1, \ldots, N \, ,$$

with factors initialized to be balanced, i.e.:

$$W_{j+1}^\top(0) W_{j+1}(0) = W_j(0) W_j^\top(0) \quad , \, j = 1, \ldots, N - 1 \, .$$

Then, the product matrix $W(t) = W_N(t) \cdots W_1(t)$ obeys the following dynamics:

$$\dot{W}(t) = -\sum\nolimits_{j=1}^{N} \left[W(t) W^\top(t) \right]^{\frac{j-1}{N}} \cdot \nabla \ell \big(W(t) \big) \cdot \left[W^\top(t) W(t) \right]^{\frac{N-j}{N}} \, ,$$

---

[13][1] and [19] also considered settings where there are multiple global minima, but in these too there was just one solution to which optimization could converge, leaving only the question of what path is taken to reach it.

[14]In addition to standard linear neural networks, [22] also analyzed "linear convolutional networks", characterized by a particular weight sharing pattern. For such models, the implicit regularization was found to promote sparsity in the frequency domain, in a manner which does depend on depth.

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

An additional result we will use is the following technical lemma:

Lemma 4. *Let $\alpha \geq \frac{1}{2}$ and $g : [0, \infty) \to \mathbb{R}$ be a continuous function. Consider the initial value problem:*

$$s(0) = s_0 \quad , \quad \dot{s}(t) = (s^2(t))^\alpha \cdot g(t) \quad \forall t \geq 0, \tag{12}$$

where $s_0 \in \mathbb{R}$. Then, as long as it does not diverge to $\pm\infty$, the solution to this problem ($s(t)$) has the same sign as its initial value (s_0). That is, $s(t)$ is identically zero if $s_0 = 0$, is positive if $s_0 > 0$, and is negative if $s_0 < 0$.

Proof. If $\alpha = 1/2$, the solution to Equation (12) is:

$$s(t) = \begin{cases} s_0 \cdot \exp\left(\int_{t'=0}^{t} g(t')dt' \right) & , s_0 > 0 \\ s_0 \cdot \exp\left(-\int_{t'=0}^{t} g(t')dt' \right) & , s_0 < 0 \\ 0 & , s_0 = 0 \end{cases}.$$

This solution does not diverge in finite time (regardless of the chosen $g(\cdot)$), and obviously preserves the sign of its initial value.

If $\alpha > 1/2$, Equation (12) is solved by:

$$s(t) = \begin{cases} \left(s_0^{-2\alpha+1} + (-2\alpha+1) \int_{t'=0}^{t} g(t')dt' \right)^{\frac{1}{-2\alpha+1}} & , s_0 > 0 \\ -\left((-s_0)^{-2\alpha+1} - (-2\alpha+1) \int_{t'=0}^{t} g(t')dt' \right)^{\frac{1}{-2\alpha+1}} & , s_0 < 0 \\ 0 & , s_0 = 0 \end{cases}.$$

In this case, divergence in finite time can take place (depending on the choice of $g(\cdot)$), but nonetheless the sign of $s(t)$ is preserved until that happens. $\square$

C Deferred proofs

C.1 Proof of Theorem 2

For convenience, throughout the proof we replace the notation $\bar{W}_{\text{deep}}$ by W_{deep}^*. We also define a linear operator $\mathcal{A}$ which specifies all m measurements:

$$\mathcal{A} : \mathbb{R}^{d,d} \to \mathbb{R}^m \quad , \quad \mathcal{A}(W) = \begin{pmatrix} \langle A_1, W \rangle \\ \vdots \\ \langle A_m, W \rangle \end{pmatrix},$$

and its adjoint operator $\mathcal{A}^\dagger$:

$$\mathcal{A}^\dagger : \mathbb{R}^m \to \mathbb{R}^{d,d} \quad , \quad \mathcal{A}^\dagger(\mathbf{r}) = \sum_{i=1}^{m} r_i A_i.$$

Then we can rewrite the loss function in Equation (2) as:

$$\ell(W) = \frac{1}{2} \|\mathcal{A}(W) - \mathbf{y}\|_2^2,$$

where $\mathbf{y} := (y_1, \ldots, y_m)^\top \in \mathbb{R}^m$. The gradient of $\ell(\cdot)$ can be expressed as:

$$\nabla \ell(W) = \mathcal{A}^\dagger(\mathcal{A}(W) - \mathbf{y}).$$

We consider a fixed $\alpha > 0$ for now, and will take the limit $\alpha \to 0^+$ later. Recall that gradient flow is run on the objective $\phi(W_1, \ldots, W_N) = \ell(W_N \cdots W_1)$, with initialization $W_j(0) = \alpha I$, $j =$

$1, \ldots, N$. From Lemma 3, we know that the product matrix $W(t) = W_N(t) \cdots W_1(t)$ evolves by:

$$\dot{W}(t) = -\sum_{j=1}^{N} \left[W(t)W^\top(t) \right]^{\frac{j-1}{N}} \cdot \nabla \ell(W(t)) \cdot \left[W^\top(t)W(t) \right]^{\frac{N-j}{N}}$$

$$= -\sum_{j=1}^{N} \left[W(t)W^\top(t) \right]^{\frac{j-1}{N}} \cdot \mathcal{A}^\dagger(\mathbf{r}(t)) \cdot \left[W^\top(t)W(t) \right]^{\frac{N-j}{N}} \quad , \ t \in \mathbb{R}_{\geq 0} \,, \qquad (13)$$

$$W(0) = \alpha^N I \,,$$

where $\mathbf{r}(t) := \mathcal{A}(W(t)) - \mathbf{y}$ is the vector of residuals at time t. Since $A_1, \ldots, A_m$ are symmetric and commutable, they are simultaneously (orthogonally) diagonalizable, *i.e.* there exists an orthogonal matrix $O \in \mathbb{R}^{d,d}$ such that $\tilde{A}_i := OA_iO^\top$, $i = 1, \ldots, m$, are all diagonal. Consider a change of variables $\tilde{W}(t) := OW(t)O^\top$, and denote $\tilde{\mathcal{A}}^\dagger(\mathbf{r}) := O\mathcal{A}^\dagger(\mathbf{r})O^\top = \sum_{i=1}^{m} r_i \tilde{A}_i$. Then it follows from Equation (13) that:

$$\dot{\tilde{W}}(t) = -\sum_{j=1}^{N} \left[\tilde{W}(t)\tilde{W}^\top(t) \right]^{\frac{j-1}{N}} \cdot \tilde{\mathcal{A}}^\dagger(\mathbf{r}(t)) \cdot \left[\tilde{W}^\top(t)\tilde{W}(t) \right]^{\frac{N-j}{N}} \quad , \ t \in \mathbb{R}_{\geq 0} \,,$$

$$\tilde{W}(0) = \alpha^N I \,. \qquad (14)$$

Notice that: *(i)* $\tilde{W}(0)$ is diagonal; and *(ii)* if $\tilde{W}(t)$ is diagonal then so is $\dot{\tilde{W}}(t)$. We may therefore set the off-diagonal elements of $\tilde{W}(t)$ to zero, and solve for the diagonal ones:

$$\dot{\tilde{W}}_{kk}(t) = -N \cdot \left(\tilde{W}_{kk}^2(t) \right)^{\frac{N-1}{N}} \cdot \tilde{\mathcal{A}}_{kk}^\dagger(\mathbf{r}(t)), \ \ \tilde{W}_{kk}(0) = \alpha^N, \ \ t \in \mathbb{R}_{\geq 0}, \ \ k = 1, \ldots, d. \quad (15)$$

By Lemma 4, $\tilde{W}_{kk}(t)$ maintains the sign of its initialization, meaning it stays positive. Moreover, since by assumption $N \geq 3$, the solution to Equation (15) is:

$$\tilde{W}_{kk}(t) = \alpha^N \left(1 + (N-2)\alpha^{N-2} \cdot \tilde{\mathcal{A}}_{kk}^\dagger(\mathbf{s}(t)) \right)^{-\frac{N}{N-2}} , \quad t \in \mathbb{R}_{\geq 0}, \quad k = 1, \ldots, d \,,$$

where $\mathbf{s}(t) := \int_{t'=0}^{t} (\mathbf{r}(t'))dt'$. The matrix $\tilde{W}(t)$ thus has positive elements on its diagonal (and zeros elsewhere), and takes the following form:

$$\tilde{W}(t) = \alpha^N \left[I_d + (N-2)\alpha^{N-2} \cdot \tilde{\mathcal{A}}^\dagger(\mathbf{s}(t)) \right]^{-\frac{N}{N-2}} , \qquad (16)$$

where I_d is the $d \times d$ identity matrix, and $[\,\cdot\,]^{-N/(N-2)}$ is a negative power operator defined over positive definite matrices. We assume $W_{\text{deep},\infty}(\alpha) := \lim_{t\to\infty} W(t)$ exists, and so we may write:

$$\tilde{W}_{\text{deep},\infty}(\alpha) := O\, W_{\text{deep},\infty}(\alpha)\, O^\top = \lim_{t\to\infty} \tilde{W}(t) = \alpha^N \left[I_d - \tilde{\mathcal{A}}^\dagger(\boldsymbol{\nu}_\infty(\alpha)) \right]^{-\frac{N}{N-2}} ,$$

where $\boldsymbol{\nu}_\infty(\alpha) := -(N-2)\alpha^{N-2} \lim_{t\to\infty} \mathbf{s}(t)$.[15] Since $\{\tilde{W}(t)\}_t$ are diagonal, $\tilde{W}_{\text{deep},\infty}(\alpha)$ is diagonal. Additionally, positive definiteness of $\{\tilde{W}(t)\}_t$ implies that $\tilde{W}_{\text{deep},\infty}(\alpha)$ is positive definite as well (it cannot have zero eigenvalues as it is given by a negative power operator). This means that:

$$I_d - \tilde{\mathcal{A}}^\dagger(\boldsymbol{\nu}_\infty(\alpha)) = \alpha^{-N} \left[\tilde{W}_{\text{deep},\infty}(\alpha) \right]^{-\frac{N-2}{N}} \succ 0 \implies \tilde{\mathcal{A}}^\dagger(\boldsymbol{\nu}_\infty(\alpha)) \prec I_d. \qquad (17)$$

Now take the limit $\alpha \to 0^+$. By assumption $W_{\text{deep}}^* := \lim_{\alpha\to 0^+} W_{\text{deep},\infty}(\alpha)$ exists, so we can write:

$$\tilde{W}_{\text{deep}}^* := O\, W_{\text{deep}}^*\, O^\top = \lim_{\alpha\to 0^+} \tilde{W}_{\text{deep},\infty}(\alpha) = \lim_{\alpha\to 0^+} \alpha^N \left[I_d - \tilde{\mathcal{A}}^\dagger(\boldsymbol{\nu}_\infty(\alpha)) \right]^{-\frac{N}{N-2}} ,$$

The fact that $\{\tilde{W}_{\text{deep},\infty}(\alpha)\}_\alpha$ are diagonal and positive definite implies that:

$$\tilde{W}_{\text{deep}}^* \succeq 0 \,. \qquad (18)$$

---

[15] Existence of $\lim_{t\to\infty} \tilde{W}(t)$ implies that $\lim_{t\to\infty} \tilde{\mathcal{A}}^\dagger(\mathbf{s}(t))$ exists (see Equation (16)), which in turn, given the linear independence of $\{\tilde{A}_i\}_{i=1}^m$, indicates that $\lim_{t\to\infty} \mathbf{s}(t)$ exists as well. Similarly to [20], the remainder of the proof treats the case where $\lim_{t\to\infty} \mathbf{s}(t)$ is finite, thereby avoiding the technical load associated with infinite coordinates.

Moreover, if the k'th element on the diagonal of $\tilde{W}^*_{\text{deep}}$ is non-zero, it must hold that:

$$\lim_{\alpha \to 0^+} \alpha^N \left(1 - \tilde{\mathcal{A}}^\dagger_{kk}\left(\boldsymbol{\nu}_\infty(\alpha)\right)\right)^{-\frac{N}{N-2}} \neq 0 \implies \lim_{\alpha \to 0^+} \tilde{\mathcal{A}}^\dagger_{kk}\left(\boldsymbol{\nu}_\infty(\alpha)\right) = 1\,,$$

from which we conclude:

$$\left\langle I_d - \tilde{\mathcal{A}}^\dagger\left(\boldsymbol{\nu}_\infty(\alpha)\right), \tilde{W}^*_{\text{deep}} \right\rangle = \sum_{k=1}^{d} \left(1 - \tilde{\mathcal{A}}^\dagger_{kk}\left(\boldsymbol{\nu}_\infty(\alpha)\right)\right) \cdot \left(\tilde{W}^*_{\text{deep}}\right)_{kk} \xrightarrow[\alpha \to 0^+]{} 0\,. \qquad (19)$$

Returning to the original variables (un-diagonalizing by the orthogonal matrix O), recall that:

$$W^*_{\text{deep}} = O^\top \tilde{W}^*_{\text{deep}} \, O\,, \qquad (20)$$

and:

$$\mathcal{A}^\dagger\left(\boldsymbol{\nu}_\infty(\alpha)\right) = O^\top \tilde{\mathcal{A}}^\dagger\left(\boldsymbol{\nu}_\infty(\alpha)\right) O\,, \qquad (21)$$

therefore the following hold:

- $W^*_{\text{deep}} \succeq 0$ (by Equations (18) and (20));
- $\mathcal{A}(W^*_{\text{deep}}) = \mathbf{y}$ (by assumption);
- $\forall \alpha > 0 : \mathcal{A}^\dagger\left(\boldsymbol{\nu}_\infty(\alpha)\right) \prec I_d$ (by Equations (17) and (21)); and
- $\lim_{\alpha \to 0^+} \left\langle I_d - \tilde{\mathcal{A}}^\dagger\left(\boldsymbol{\nu}_\infty(\alpha)\right), \tilde{W}^*_{\text{deep}} \right\rangle = 0$ (by Equations (19), (20) and (21)).

Lemma 5 below then concludes the proof. $\qquad \square$

Lemma 5. *Suppose that $W^* \in \mathcal{S}^d_+$ satisfies $\mathcal{A}(W^*) = \mathbf{y}$, and that there exists a sequence of vectors $\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \ldots \in \mathbb{R}^m$ such that $\mathcal{A}^\dagger(\boldsymbol{\nu}_n) \preceq I$ for all n and $\lim_{n \to \infty}\langle I - \mathcal{A}^\dagger(\boldsymbol{\nu}_n), W^*\rangle = 0$. Then $W^* \in \operatorname{argmin}_{W \in \mathcal{S}^d_+, \mathcal{A}(W)=\mathbf{y}} \|W\|_*$.*

Proof. Recall that $\mathcal{S}^d_+$ stands for the set of (symmetric and) positive semidefinite matrices in $\mathbb{R}^{d,d}$, and $\|\cdot\|_*$ denotes matrix nuclear norm. The minimization problem being considered can be framed as a semidefinite program:[16]

$$\begin{aligned}
\text{minimize} \quad & \langle I, W \rangle \\
\text{subject to} \quad & \mathcal{A}(W) = \mathbf{y} \\
& W \in \mathcal{S}^d_+\,.
\end{aligned} \qquad (22)$$

A corresponding dual program is:

$$\begin{aligned}
\text{maximize} \quad & \boldsymbol{\nu}^\top \mathbf{y} \\
\text{subject to} \quad & \mathcal{A}^*(\boldsymbol{\nu}) \preceq I \\
& \boldsymbol{\nu} \in \mathbb{R}^m\,.
\end{aligned} \qquad (23)$$

Let OPT be the optimal value for the primal program (Equation (22)):

$$\mathsf{OPT} := \min_{W \in \mathcal{S}^d_+, \mathcal{A}(W)=\mathbf{y}} \|W\|_*\,.$$

By duality theory, for any $\boldsymbol{\nu}$ feasible in the dual program (Equation (23)), we have $\boldsymbol{\nu}^\top \mathbf{y} \leq \mathsf{OPT}$. Since W^* is feasible in the primal, and each $\boldsymbol{\nu}_n$ is feasible in the dual, it holds that:

$$0 \leq \|W^*\|_* - \mathsf{OPT} \leq \|W^*\|_* - \boldsymbol{\nu}_n^\top \mathbf{y} = \langle I, W^* \rangle - \boldsymbol{\nu}_n^\top \mathcal{A}(W^*) = \langle I - \mathcal{A}^\dagger(\boldsymbol{\nu}_n), W^* \rangle\,.$$

Taking the limit $n \to \infty$, the right hand side above becomes 0, which implies $\|W^*\|_* = \mathsf{OPT}$. $\quad \square$

---

[16]Note that for $W \in \mathcal{S}^d_+$ we have $\|W\|_* = \langle I, W \rangle$.

C.2 Proof of Proposition 1

We will choose $A_1, \ldots, A_m$ to be diagonal. This of course ensures symmetry and commutativity. Additionally, by the proof of Theorem 2 (Appendix C.1), it implies that $\bar{W}_{\text{deep}}$ is diagonal and positive semidefinite.[17] We set $A_1, \ldots, A_m$ and $y_1, \ldots, y_m$ such that the linear equations $\langle A_i, W \rangle = y_i$, $i = 1, \ldots, m$, are the following:

$$\begin{aligned} W_{11} &= W_{22} \\ W_{11} &= W_{kk} + 1 \quad, \ k = 3, 4, \ldots, d\,. \end{aligned} \tag{24}$$

Note that the matrices $A_1, \ldots, A_m$ which naturally induce these equations are (diagonal and) linearly independent, as required. We know that $\bar{W}_{\text{deep}}$ is diagonal and has minimal nuclear norm among all positive semidefinite matrices that satisfy the equations. Using the fact that for positive semidefinite matrices nuclear norm is the same as trace, one readily sees that:

$$\bar{W}_{\text{deep}} = \text{diag}(1, 1, 0, 0, \ldots, 0)\,.$$

We complete the proof by showing that in any neighborhood of $\bar{W}_{\text{deep}}$, there exists a positive semidefinite matrix that meets Equation (24) and has strictly smaller Schatten-p quasi-norm for any $0 < p < 1$. Indeed, for $\epsilon \in (0, 1)$ define:

$$\hat{W}_\epsilon := \begin{pmatrix} 1 & \epsilon & 0 & \cdots & 0 \\ \epsilon & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix} \in \mathbb{R}^{d,d}\,.$$

$\hat{W}_\epsilon$ obviously satisfies Equation (24). Additionally, it is symmetric with eigenvalues:

$$\lambda_1 = 1 + \epsilon \ , \ \ \lambda_2 = 1 - \epsilon \ , \ \ \lambda_3 = \cdots = \lambda_d = 0\,,$$

and therefore is positive semidefinite. For any $0 < p < 1$:

$$\|\hat{W}_\epsilon\|_{S_p}^p = (1 - \epsilon)^p + (1 + \epsilon)^p < 2 \cdot \left(\tfrac{1}{2}(1 + \epsilon) + \tfrac{1}{2}(1 - \epsilon)\right)^p = 2 = \|\bar{W}_{\text{deep}}\|_{S_p}^p\,,$$

where the inequality follows from $\theta_p : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, $\theta_p(x) = x^p$, being strictly concave. Noting that taking $\epsilon \to 0^+$ makes $\hat{W}_\epsilon$ arbitrarily close to $\bar{W}_{\text{deep}}$, we conclude the proof. $\qquad\square$

C.3 Proof of Lemma 1

By Theorem 1 in [7], it suffices to show that the product matrix $W(t)$ is an analytic function of t. Analytic functions are closed under summation, multiplication and composition, so the analyticity of $\ell(\cdot)$ implies that $\phi(\cdot)$ (Equation (3)) is analytic as well. It then follows (see Theorem 1.1 in [27]) that under gradient flow (Equation (4)), the factors $W_1(t), \ldots, W_N(t)$ are analytic functions of t. Therefore $W(t)$ (Equation (1)) is also analytic in t. $\qquad\square$

C.4 Proof of Theorem 3

Differentiate the analytic singular value decomposition (Equation (6)) with respect to time:

$$\dot{W}(t) = \dot{U}(t)S(t)V^\top(t) + U(t)\dot{S}(t)V^\top(t) + U(t)S(t)\dot{V}^\top(t)\,,$$

then multiply from the left by $U^\top(t)$ and from the right by $V(t)$:

$$U^\top(t)\dot{W}(t)V(t) = U^\top(t)\dot{U}(t)S(t) + \dot{S}(t) + S(t)\dot{V}^\top(t)V(t)\,,$$

where we used the fact that $U(t)$ and $V(t)$ have orthonormal columns. Restricting our attention to the diagonal elements of this matrix equation, we have:

$$\mathbf{u}_r^\top(t)\dot{W}(t)\mathbf{v}_r(t) = \langle \mathbf{u}_r(t), \dot{\mathbf{u}}_r(t)\rangle \, \sigma_r(t) + \dot{\sigma}_r(t) + \sigma_r(t) \, \langle \dot{\mathbf{v}}_r(t), \mathbf{v}_r(t)\rangle \quad, \ r = 1, \ldots, \min\{d, d'\}\,.$$

---

[17]In the proof of Theorem 2 (Appendix C.1), diagonality of $A_1, \ldots, A_m$ corresponds to the case where O — the diagonalizing matrix — is simply the identity, and therefore $\bar{W}_{\text{deep}}$ is equal to $\tilde{W}_{\text{deep}}^*$, implying that the former is indeed diagonal and positive semidefinite.

Since $\mathbf{u}_r(t)$ has constant (unit) length it holds that $\langle \mathbf{u}_r(t), \dot{\mathbf{u}}_r(t) \rangle = \frac{1}{2} \cdot \frac{d}{dt} \|\mathbf{u}_r(t)\|_2^2 = 0$, and similarly $\langle \dot{\mathbf{v}}_r(t), \mathbf{v}_r(t) \rangle = 0$. The latter equation thus simplifies to:

$$\dot{\sigma}_r(t) = \mathbf{u}_r^\top(t) \dot{W}(t) \mathbf{v}_r(t) \quad , \quad r = 1, \ldots, \min\{d, d'\} \,. \tag{25}$$

Lemma 3 from Appendix B provides the following expression for $\dot{W}(t)$:

$$\dot{W}(t) = -\sum_{j=1}^{N} \left[W(t) W^\top(t) \right]^{\frac{j-1}{N}} \cdot \nabla \ell\big(W(t)\big) \cdot \left[W^\top(t) W(t) \right]^{\frac{N-j}{N}} \,,$$

where $[\cdot]^\alpha$, $\alpha \in \mathbb{R}_{\geq 0}$, stands for a power operator defined over positive semidefinite matrices (with $\alpha = 0$ yielding identity by definition). Plugging in the analytic singular value decomposition (Equation (6)) gives:

$$\begin{aligned}
\dot{W}(t) &= -\nabla \ell\big(W(t)\big) \cdot V(t) \big(S^2(t)\big)^{\frac{N-1}{N}} V^\top(t) \\
&\quad -\sum_{j=2}^{N-1} U(t) \big(S^2(t)\big)^{\frac{j-1}{N}} U^\top(t) \cdot \nabla \ell\big(W(t)\big) \cdot V(t) \big(S^2(t)\big)^{\frac{N-j}{N}} V^\top(t) \\
&\quad -U(t) \big(S^2(t)\big)^{\frac{N-1}{N}} U^\top(t) \cdot \nabla \ell\big(W(t)\big) \,.
\end{aligned}$$

Left-multiplying by $\mathbf{u}_r^\top(t)$, right-multiplying by $\mathbf{v}_r(t)$, and using the fact that $\{\mathbf{u}_r(t)\}_r$ (columns of $U(t)$) and $\{\mathbf{v}_r(t)\}_r$ (columns of $V(t)$) are orthonormal sets, we obtain:

$$\begin{aligned}
\mathbf{u}_r^\top(t) \dot{W}(t) \mathbf{v}_r(t) &= -\mathbf{u}_r^\top(t) \nabla \ell\big(W(t)\big) \mathbf{v}_r(t) \cdot (\sigma_r^2(t))^{\frac{N-1}{N}} \\
&\quad -\sum_{j=2}^{N-1} (\sigma_r^2(t))^{\frac{j-1}{N}} \cdot \mathbf{u}_r^\top(t) \nabla \ell\big(W(t)\big) \mathbf{v}_r(t) \cdot (\sigma_r^2(t))^{\frac{N-j}{N}} \\
&\quad -(\sigma_r^2(t))^{\frac{N-1}{N}} \cdot \mathbf{u}_r^\top(t) \nabla \ell\big(W(t)\big) \mathbf{v}_r(t) \\
&= -N \cdot (\sigma_r^2(t))^{\frac{N-1}{N}} \cdot \mathbf{u}_r^\top(t) \nabla \ell\big(W(t)\big) \mathbf{v}_r(t) \,.
\end{aligned}$$

Combining this with Equation (25) yields the sought-after Equation (7).

To complete the proof, it remains to show that if the matrix factorization is non-degenerate (has depth $N \geq 2$), singular values need not be signed, *i.e.* we may assume $\sigma_r(t) \geq 0$ for all t. Equation (7), along with Lemma 4, imply that if $N \geq 2$, $\sigma_r(t)$ will never switch sign. Therefore, either $\sigma_r(t) \geq 0$ for all t, or alternatively, this will hold if we take away a minus sign from $\sigma_r(t)$ and absorb it into $\mathbf{u}_r(t)$ (or $\mathbf{v}_r(t)$). $\qquad \square$

C.5 Proof of Lemma 2

A real analytic function is either identically zero, or admits a zero set with no accumulation points (*cf.* [30]). For any $r \in \{1, \ldots, \min\{d, d'\}\}$, applying this fact to the signed singular value $\sigma_r(t)$, while taking into account our assumption of it being different from zero at initialization, we conclude that the set of times t for which it vanishes has no accumulation points. Similarly, for any $r, r' \in \{1, \ldots, \min\{d, d'\}\}$, $r \neq r'$, we assumed that $\sigma_r^2(t) - \sigma_{r'}^2(t)$ is different from zero at initialization, and thus the set of times t for which it vanishes is free from accumulation points. Overall, any time t for which $\sigma_r(t) = 0$ for some r, or $\sigma_r^2(t) = \sigma_{r'}^2(t)$ for some $r \neq r'$, must be isolated, *i.e.* surrounded by a neighborhood in which none of these conditions are met. Accordingly, hereafter, we assume $\forall r : \sigma_r(t) \neq 0$ and $\forall r \neq r' : \sigma_r^2(t) \neq \sigma_{r'}^2(t)$, knowing that for times t in which this does not hold, $\dot{U}(t)$ and $\dot{V}(t)$ can be inferred by continuity.

We now follow a series of steps adopted from [46], to derive expressions for $\dot{U}(t)$ and $\dot{V}(t)$ in terms of $U(t)$, $V(t)$, $S(t)$ and $\dot{W}(t)$. Differentiate the analytic singular value decomposition (Equation (6)) with respect to time:

$$\dot{W}(t) = \dot{U}(t) S(t) V^\top(t) + U(t) \dot{S}(t) V^\top(t) + U(t) S(t) \dot{V}^\top(t) \,. \tag{26}$$

Multiplying from the left by $U^\top(t)$ and from the right by $V(t)$, we have:

$$U^\top(t) \dot{W}(t) V(t) = U^\top(t) \dot{U}(t) S(t) + \dot{S}(t) + S(t) \dot{V}^\top(t) V(t) \,, \tag{27}$$

where we used the fact that $U(t)$ and $V(t)$ have orthonormal columns. This orthonormality also implies that $U^\top(t)\dot{U}(t)$ and $\dot{V}^\top(t)V(t)$ are skew-symmetric,[18] and in particular have zero diagonals. Since $S(t)$ is diagonal, $U^\top(t)\dot{U}(t)S(t)$ and $S(t)\dot{V}^\top(t)V(t)$ have zero diagonals as well. On the other hand $\dot{S}(t)$ holds zeros outside its diagonal, and so we may write:

$$\bar{I}_{\min\{d,d'\}} \odot (U^\top(t)\dot{W}(t)V(t)) = U^\top(t)\dot{U}(t)S(t) + S(t)\dot{V}^\top(t)V(t)\,, \tag{28}$$

where $\odot$ stands for Hadamard (element-wise) product, and $\bar{I}_{\min\{d,d'\}}$ is a $\min\{d,d'\} \times \min\{d,d'\}$ matrix holding zeros on its diagonal and ones elsewhere. Taking transpose of Equation (28), while recalling that $U^\top(t)\dot{U}(t)$ and $\dot{V}^\top(t)V(t)$ are skew-symmetric, we have:

$$\bar{I}_{\min\{d,d'\}} \odot (V^\top(t)\dot{W}^\top(t)U(t)) = -S(t)U^\top(t)\dot{U}(t) - \dot{V}^\top(t)V(t)S(t)\,. \tag{29}$$

Right-multiply Equation (28) by $S(t)$, left-multiply Equation (29) by $S(t)$, and add:

$$\bar{I}_{\min\{d,d'\}} \odot (U^\top(t)\dot{W}(t)V(t)S(t) + S(t)V^\top(t)\dot{W}^\top(t)U(t)) = U^\top(t)\dot{U}(t)S^2(t) - S^2(t)U^\top(t)\dot{U}(t)\,.$$

Since we assume diagonal elements of $S^2(t)$ are distinct ($\sigma_r^2(t) \neq \sigma_{r'}^2(t)$ for $r \neq r'$), this implies:

$$U^\top(t)\dot{U}(t) = H(t) \odot \left[U^\top(t)\dot{W}(t)V(t)S(t) + S(t)V^\top(t)\dot{W}^\top(t)U(t)\right]\,,$$

where the matrix $H(t) \in \mathbb{R}^{\min\{d,d'\},\min\{d,d'\}}$ is defined by:

$$H_{r,r'}(t) := \begin{cases} \left(\sigma_{r'}^2(t) - \sigma_r^2(t)\right)^{-1} & , r \neq r' \\ 0 & , r = r' \end{cases}\,. \tag{30}$$

Multiplying from the left by $U(t)$ yields:

$$P_{U(t)}\dot{U}(t) = U(t)\left(H(t) \odot \left[U^\top(t)\dot{W}(t)V(t)S(t) + S(t)V^\top(t)\dot{W}^\top(t)U(t)\right]\right)\,, \tag{31}$$

with $P_{U(t)} := U(t)U^\top(t)$ being the projection onto the subspace spanned by the (orthonormal) columns of $U(t)$. Denote by $P_{U_\perp(t)}$ the projection onto the orthogonal complement, *i.e.* $P_{U_\perp(t)} := I_d - U(t)U^\top(t)$, where I_d is the $d \times d$ identity matrix. Apply $P_{U_\perp(t)}$ to both sides of Equation (26):

$$P_{U_\perp(t)}\dot{W}(t) = P_{U_\perp(t)}\dot{U}(t)S(t)V^\top(t) + P_{U_\perp(t)}U(t)\dot{S}(t)V^\top(t) + P_{U_\perp(t)}U(t)S(t)\dot{V}^\top(t)\,.$$

Note that $P_{U_\perp(t)}U(t) = 0$, and multiply from the right by $V(t)S^{-1}(t)$ (the latter is well-defined since we assume diagonal elements of $S(t)$ are non-zero — $\sigma_r(t) \neq 0$):

$$P_{U_\perp(t)}\dot{U}(t) = P_{U_\perp(t)}\dot{W}(t)V(t)S^{-1}(t) = \left(I_d - U(t)U^\top(t)\right)\dot{W}(t)V(t)S^{-1}(t)\,. \tag{32}$$

Adding Equations (31) and (32), we obtain an expression for $\dot{U}(t)$:

$$\begin{aligned}
\dot{U}(t) &= P_{U(t)}\dot{U}(t) + P_{U_\perp(t)}\dot{U}(t) \\
&= U(t)\left(H(t) \odot \left[U^\top(t)\dot{W}(t)V(t)S(t) + S(t)V^\top(t)\dot{W}^\top(t)U(t)\right]\right) \\
&\quad + \left(I_d - U(t)U^\top(t)\right)\dot{W}(t)V(t)S^{-1}(t)\,.
\end{aligned} \tag{33}$$

By returning to Equations (28) and (29), switching the directions from which they were multiplied by $S(t)$ (*i.e.* multiplying Equation (28) from the left and Equation (29) from the right), and continuing similarly to above, an analogous expression for $\dot{V}(t)$ is derived:

$$\begin{aligned}
\dot{V}(t) &= V(t)\left(H(t) \odot \left[S(t)U^\top(t)\dot{W}(t)V(t) + V^\top(t)\dot{W}^\top(t)U(t)S(t)\right]\right) \\
&\quad + \left(I_{d'} - V(t)V^\top(t)\right)\dot{W}^\top(t)U(t)S^{-1}(t)\,,
\end{aligned} \tag{34}$$

where $I_{d'}$ is the $d' \times d'$ identity matrix.

Next, we invoke Lemma 3 from Appendix B, which provides an expression for $\dot{W}(t)$:

$$\dot{W}(t) = -\sum_{j=1}^{N} \left[W(t)W^\top(t)\right]^{\frac{j-1}{N}} \cdot \nabla\ell\big(W(t)\big) \cdot \left[W^\top(t)W(t)\right]^{\frac{N-j}{N}}\,, \tag{35}$$

---

[18]To see this, note that $U^\top(t)U(t)$ is constant, thus its derivative with respect to time is equal to zero, *i.e.* $\dot{U}^\top(t)U(t) + U^\top(t)\dot{U}(t) = 0$ (by an analogous argument $\dot{V}^\top(t)V(t) + V^\top(t)\dot{V}(t) = 0$ holds as well).

where $[\,\cdot\,]^{\alpha}$, $\alpha \in \mathbb{R}_{\geq 0}$, stands for a power operator defined over positive semidefinite matrices (with $\alpha = 0$ yielding identity by definition). Plug the analytic singular value decomposition (Equation (6)) into Equation (35):

$$
\begin{aligned}
\dot{W}(t) \;=\; & -\nabla\ell\big(W(t)\big) \cdot V(t)\big(S^2(t)\big)^{\frac{N-1}{N}} V^{\top}(t) \\
& -\sum\nolimits_{j=2}^{N-1} U(t)\big(S^2(t)\big)^{\frac{j-1}{N}} U^{\top}(t) \cdot \nabla\ell\big(W(t)\big) \cdot V(t)\big(S^2(t)\big)^{\frac{N-j}{N}} V^{\top}(t) \\
& -U(t)\big(S^2(t)\big)^{\frac{N-1}{N}} U^{\top}(t) \cdot \nabla\ell\big(W(t)\big)\,.
\end{aligned}
\tag{36}
$$

From this it follows that:

$$
\begin{aligned}
U^{\top}(t)\dot{W}(t)V(t) \;=\; & -U^{\top}(t)\nabla\ell\big(W(t)\big)V(t)\big(S^2(t)\big)^{\frac{N-1}{N}} \\
& -\sum\nolimits_{j=2}^{N-1} \big(S^2(t)\big)^{\frac{j-1}{N}} U^{\top}(t)\nabla\ell\big(W(t)\big)V(t)\big(S^2(t)\big)^{\frac{N-j}{N}} \\
& -\big(S^2(t)\big)^{\frac{N-1}{N}} U^{\top}(t)\nabla\ell\big(W(t)\big)V(t) \\
\;=\; & -G(t) \odot \big[U^{\top}(t)\nabla\ell\big(W(t)\big)V(t)\big]\,,
\end{aligned}
\tag{37}
$$

where $G(t) \in \mathbb{R}^{\min\{d,d'\},\min\{d,d'\}}$ is defined by:

$$
G_{r,r'}(t) := \sum\nolimits_{j=1}^{N} (\sigma_r^2(t))^{\frac{j-1}{N}} (\sigma_{r'}^2(t))^{\frac{N-j}{N}}\,.
\tag{38}
$$

Since $G(t)$ is symmetric (and $S(t)$ is diagonal), Equation (37) implies:

$$
\begin{aligned}
& U^{\top}(t)\dot{W}(t)V(t)S(t) + S(t)V^{\top}(t)\dot{W}^{\top}(t)U(t) \\
& \quad = -G(t) \odot \big[U^{\top}(t)\nabla\ell\big(W(t)\big)V(t)S(t) + S(t)V^{\top}(t)\nabla\ell^{\top}\big(W(t)\big)U(t)\big]\,.
\end{aligned}
$$

Taking Hadamard product by $H(t)$ (Equation (30)) we obtain:

$$
\begin{aligned}
& H(t) \odot \big[U^{\top}(t)\dot{W}(t)V(t)S(t) + S(t)V^{\top}(t)\dot{W}^{\top}(t)U(t)\big] \\
& \quad = -F(t) \odot \big[U^{\top}(t)\nabla\ell\big(W(t)\big)V(t)S(t) + S(t)V^{\top}(t)\nabla\ell^{\top}\big(W(t)\big)U(t)\big]\,,
\end{aligned}
$$

where $F(t) := H(t) \odot G(t)$ is given by:

$$
F_{r,r'}(t) := \begin{cases} \big((\sigma_{r'}^2(t))^{1/N} - (\sigma_r^2(t))^{1/N}\big)^{-1} & ,\, r \neq r' \\ 0 & ,\, r = r' \end{cases}\,.
\tag{39}
$$

Plug this into Equation (33):

$$
\begin{aligned}
\dot{U}(t) \;=\; & -U(t)\big(F(t) \odot \big[U^{\top}(t)\nabla\ell\big(W(t)\big)V(t)S(t) + S(t)V^{\top}(t)\nabla\ell^{\top}\big(W(t)\big)U(t)\big]\big) \\
& +\big(I_d - U(t)U^{\top}(t)\big)\dot{W}(t)V(t)S^{-1}(t)\,.
\end{aligned}
\tag{40}
$$

The first term on the right-hand side here complies with the result we seek to prove (Equation (8)). To treat the second term, we again invoke Equation (36), noting that the matrix $P_{U_{\perp}(t)} := I_d - U(t)U^{\top}(t)$ (projection onto the orthogonal complement of the subspace spanned by the columns of $U(t)$) produces zero when right-multiplied by $U(t)$. This implies:

$$
\big(I_d - U(t)U^{\top}(t)\big)\dot{W}(t)V(t)S^{-1}(t) = -\big(I_d - U(t)U^{\top}(t)\big)\nabla\ell\big(W(t)\big)V(t)\big(S^2(t)\big)^{\frac{1}{2}-\frac{1}{N}}\,.
$$

Plugging this back into Equation (40) yields Equation (8) — sought-after result. The analogous Equation (9) can be derived in a similar fashion (by incorporating Equation (35) into Equation (34), as we have done for Equation (33)). $\qquad\square$

C.6   Proof of Corollary 1

As stated in the proof of Lemma 2 (Appendix C.5), for all t but a set of isolated points it holds that $\forall r : \sigma_r(t) \neq 0$ and $\forall r \neq r' : \sigma_r^2(t) \neq \sigma_{r'}^2(t)$, meaning Equations (8) and (9) are well-defined. We

will initially assume this to be the case, and then treat isolated points by taking limits. Left-multiply Equation (8) by $U^\top(t)$ and Equation (9) by $V^\top(t)$:

$$U^\top(t)\dot{U}(t) = -F(t) \odot \left[U^\top(t)\nabla\ell(W(t))V(t)S(t) + S(t)V^\top(t)\nabla\ell^\top(W(t))U(t)\right]$$
$$V^\top(t)\dot{V}(t) = -F(t) \odot \left[S(t)U^\top(t)\nabla\ell(W(t))V(t) + V^\top(t)\nabla\ell^\top(W(t))U(t)S(t)\right],$$

where we have used the fact that $U(t)$ and $V(t)$ have orthonormal columns. Right-multiplying the first equation by $S(t)$, left-multiplying the second by $S(t)$, and then subtracting, we obtain:

$$U^\top(t)\dot{U}(t)S(t) - S(t)V^\top(t)\dot{V}(t)$$
$$= -F(t) \odot \left[U^\top(t)\nabla\ell(W(t))V(t)S^2(t) - S^2(t)U^\top(t)\nabla\ell(W(t))V(t)\right]$$
$$= -F(t) \odot E(t) \odot \left[U^\top(t)\nabla\ell(W(t))V(t)\right],$$

where the matrix $E(t) \in \mathbb{R}^{\min\{d,d'\},\min\{d,d'\}}$ is defined by: $E_{r,r'}(t) := \sigma_{r'}^2(t) - \sigma_r^2(t)$. Recalling the definition of $F(t)$ (Equation (39)), we have:

$$U^\top(t)\dot{U}(t)S(t) - S(t)V^\top(t)\dot{V}(t) = -\bar{I}_{\min\{d,d'\}} \odot G(t) \odot \left[U^\top(t)\nabla\ell(W(t))V(t)\right], \quad (41)$$

where $G(t) \in \mathbb{R}^{\min\{d,d'\},\min\{d,d'\}}$ is the matrix defined in Equation (38), and $\bar{I}_{\min\{d,d'\}}$ is a matrix of the same size, with zeros on its diagonal and ones elsewhere. Since by assumption $\forall r : \sigma_r(t) \neq 0$, the matrix $G(t)$ does not contain zero elements. Therefore when $\dot{U}(t) = 0$ and $\dot{V}(t) = 0$, leading the left-hand side of Equation (41) to vanish, it must be that $U^\top(t)\nabla\ell(W(t))V(t)$ is diagonal.

To complete the proof, it remains to treat those isolated times t for which the conditions $\forall r : \sigma_r(t) \neq 0$ and $\forall r \neq r' : \sigma_r^2(t) \neq \sigma_{r'}^2(t)$ do not all hold, and thus our derivation of Equation (41) may be invalid. Since both sides of the equation are continuous, it carries over to such isolated times, and is in fact applicable to any t. Accordingly, any t for which $\dot{U}(t) = 0$ and $\dot{V}(t) = 0$ admits $\bar{I}_{\min\{d,d'\}} \odot G(t) \odot \left[U^\top(t)\nabla\ell(W(t))V(t)\right] = 0$. Recalling the definition of $G(t)$ (Equation (38)), it is clear that the latter equality implies diagonality of $U^\top(t)\nabla\ell(W(t))V(t)$ if $\forall r : \sigma_r(t) \neq 0$. This means that the sought-after result holds if $\forall r : \sigma_r(t) \neq 0$ for every t. Recollecting that the product matrix is initialized to be full-rank ($\forall r : \sigma_r(0) \neq 0$), and invoking our assumption on the factorization being non-degenerate ($N \geq 2$), we apply Lemma 4 (from Appendix C.4) to the evolution of $\{\sigma_r(t)\}_r$ (Equation (7) in Theorem 3) and conclude the proof. $\qquad\square$

D Extension of [20] to asymmetric matrix factorization

Extending Theorem 1 from [20] to asymmetric (depth-2) matrix factorizations boils down to proving the following proposition:

Proposition 2. *Consider gradient flow on the objective:*

$$\phi(W_1, W_2) = \ell(W_2 W_1) = \frac{1}{2}\sum_{i=1}^m (y_i - \langle A_i, W_2 W_1\rangle)^2,$$

with $W_1, W_2 \in \mathbb{R}^{d,d}$ initialized to αI, $\alpha > 0$, and denote by $W_{\mathrm{sha},\infty}(\alpha)$ the product matrix obtained at the end of optimization (i.e. $W_{\mathrm{sha},\infty}(\alpha) := \lim_{t\to\infty} W_2(t)W_1(t)$ where $W_j(0) = \alpha I$ and $\dot{W}_j(t) = -\frac{\partial\phi}{\partial W_j}(W_1(t), W_2(t))$ for $t \in \mathbb{R}_{\geq 0}$). Assume the measurement matrices $A_1, \ldots, A_m$ commute. Then, if $\bar{W}_{\mathrm{sha}} := \lim_{\alpha\to 0} W_{\mathrm{sha},\infty}(\alpha)$ exists and is a global optimum for Equation (2) with $\ell(\bar{W}_{\mathrm{sha}}) = 0$, it holds that $\bar{W}_{\mathrm{sha}} \in \mathrm{argmin}_{W\in\mathcal{S}_+^d, \ell(W)=0} \|W\|_*$, i.e. $\bar{W}_{\mathrm{sha}}$ is a global optimum with minimal nuclear norm.

Proof. We follow the proof of Theorem 2 (Appendix C.1) up until Equation (15). Equation (13), specialized to $N = 2$, yields dynamics for the product matrix $W(t) = W_2(t)W_1(t)$:

$$\dot{W}(t) = -\mathcal{A}^*(\mathbf{r}(t)) \cdot \left[W^\top(t)W(t)\right]^{\frac{1}{2}} - \left[W(t)W^\top(t)\right]^{\frac{1}{2}} \cdot \mathcal{A}^*(\mathbf{r}(t)),$$
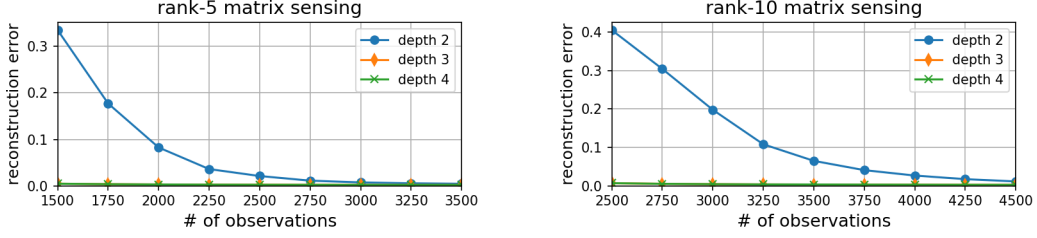$$W(0) = \alpha^2 I.$$

$$(42)$$

Figure 4: Matrix sensing via gradient descent over deep matrix factorizations. This figure is identical to Figure 1, except that reconstruction of a ground truth matrix is based not on a randomly chosen subset of entries, but on a set of random projections (*i.e.* on $\{\langle A_i, W^* \rangle\}_{i=1}^m$ where W^* is the ground truth and $A_1, \ldots, A_m$ are measurement matrices drawn independently from a Gaussian distribution). For further details on this experiment see Appendix E.2.
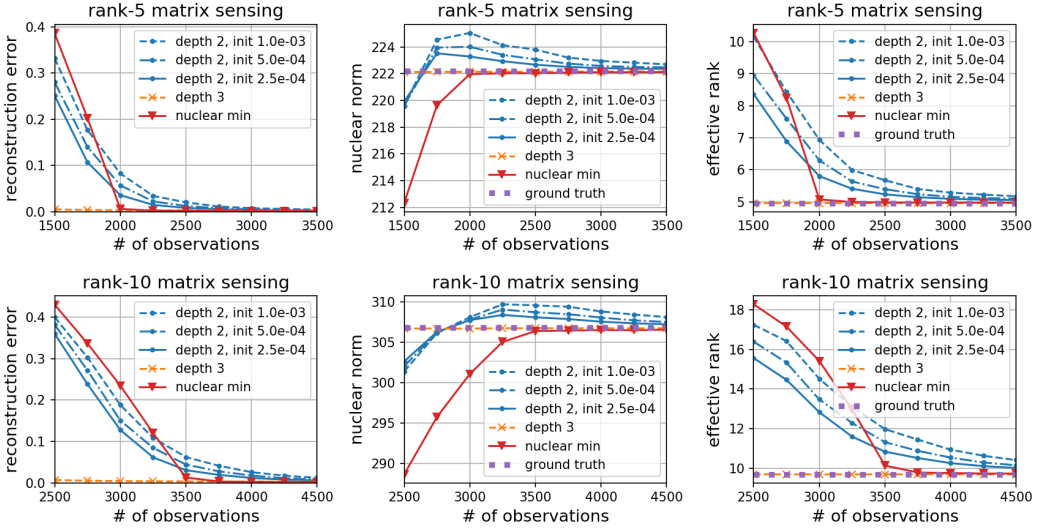

Figure 5: Evaluation of nuclear norm as the implicit regularization in deep matrix factorization on matrix sensing tasks. This figure is identical to Figure 2, except that reconstruction of a ground truth matrix is based not on a randomly chosen subset of entries, but on a set of random projections (*i.e.* on $\{\langle A_i, W^* \rangle\}_{i=1}^m$ where W^* is the ground truth and $A_1, \ldots, A_m$ are measurement matrices drawn independently from a Gaussian distribution). For further details on this experiment see Appendix E.2.

Equation (15), along with Lemma 4, imply that $\tilde{W}_{kk}(t)$ maintains the sign of its initialization, *i.e.* is positive. The diagonal matrix $\tilde{W}(t)$ is therefore positive definite, and so is the product matrix $W(t) = O^\top \tilde{W}(t) O$. Equation (42) thus becomes:

$$\dot{W}(t) = -\mathcal{A}^*(\mathbf{r}(t)) \cdot W(t) - W(t) \cdot \mathcal{A}^*(\mathbf{r}(t)),$$
$$W(0) = \alpha^2 I. \tag{43}$$

The dynamics in Equation (43) are precisely those developed in [20] for a symmetric matrix factorization. The proof of Theorem 1 there can now be applied as is, establishing the desired result. $\qquad\square$

E   Further experiments and implementation details

E.1   Further experiments

Figures 4, 5 and 6 present matrix sensing experiments supplementing the matrix completion experiments reported in Figures 1, 2 and 3 respectively.
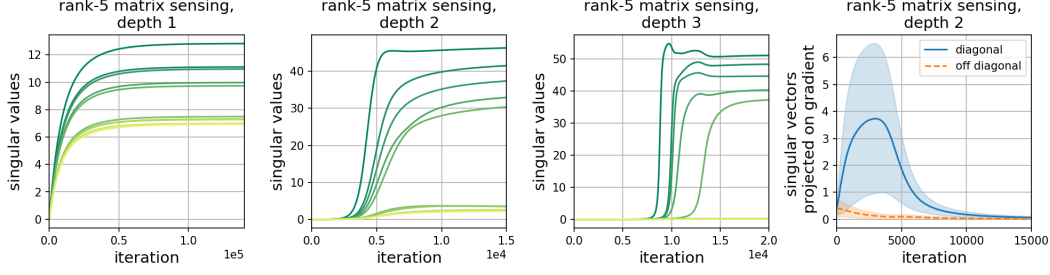
Figure 6: Dynamics of gradient descent over deep matrix factorizations on a matrix sensing task. This figure is identical to the top row of Figure 3, except that training is based not on 2000 randomly chosen entries of the ground truth matrix, but on 2000 random projections (*i.e.* on $\{\langle A_i, W^* \rangle\}_{i=1}^{2000}$ where W^* is the ground truth and $A_1, \ldots, A_{2000}$ are measurement matrices drawn independently from a Gaussian distribution). For further details on this experiment see Appendix E.2.

E.2 Implementation details

In this appendix we provide implementation details omitted from the descriptions of our experiments (Figures 1, 2, 3, 4, 5 and 6). Our implementation is based on Python, with PyTorch ([40]) for realizing deep matrix factorizations and CVXPY ([13, 2]) for finding minimum nuclear norm solutions. Source code for reproducing our results can be found in `https://github.com/roosephu/deep_matrix_factorization`.

When referring to a random rank-r matrix with size $d \times d'$, we mean a product $UV^\top$, where the entries of $U \in \mathbb{R}^{d,r}$ and $V \in \mathbb{R}^{d',r}$ are drawn independently from the standard normal distribution. Randomly chosen observed entries in synthetic matrix completion tasks (Figures 1, 2 and top row of Figure 3) were selected uniformly (without repetition). In synthetic matrix sensing tasks (Figures 4, 5 and 6), entries of all measurement (projection) matrices were drawn independently from the standard normal distribution. When varying the number of observations in synthetic matrix completion and sensing (Figures 1, 2, 4 and 5), we evaluated increments of 250. Training on MovieLens 100K dataset (bottom row of Figure 3) comprised fitting 10000 randomly (uniformly) chosen samples from the 100000 entries given in the 943×1682 user-movie rating matrix (see [24]).

In all experiments, deep matrix factorizations were trained by (full batch) gradient descent applied to ℓ_2 loss over the observed entries (in matrix completion tasks) or given projections (in matrix sensing tasks), with no explicit regularization. Gradient descent was initialized by independently sampling all weights from a Gaussian distribution with zero mean and configurable standard deviation. Learning rates were fixed throughout optimization, and the stopping criterion was training loss reaching value lower than 10^{-6} (or 10^6 iterations elapsing). In the nuclear norm evaluation experiments (Figures 2 and 5), learning rate and standard deviation of initialization for gradient descent were assigned values from the set $\{10^{-3}, 5 \cdot 10^{-4}, 2.5 \cdot 10^{-4}\}$. In the dynamics illustration experiments (Figures 3 and 6), displayed results correspond to both learning rate and standard deviation for initialization being 10^{-3}.

In figures 1, 2, 4 and 5, each error bar marks standard deviation of the respective result over three trials differing in random seed for initialization of gradient descent. Reconstruction error with respect to a ground truth matrix W^* is based on normalized Frobenius distance, *i.e.* for a solution W it is $\|W - W^*\|_F / \|W^*\|_F$. In experiments with matrix completion and sensing under varying number of observations (Figures 1, 2, 4 and 5), plots begin at the smallest number for which stable results were obtained, and end when all evaluated methods are close to zero reconstruction error. For the dynamics illustration experiments (Figures 3 and 6), plots showing singular values hold 10 curves corresponding to the largest ones.