Implicit Regularization in Tensor Factorization

Noam Razin 1 * Asaf Maman 1 * Nadav Cohen 1

Abstract

Recent efforts to unravel the mystery of implicit regularization in deep learning have led to a theoretical focus on matrix factorization — matrix completion via linear neural network. As a step further towards practical deep learning, we provide the first theoretical analysis of implicit regularization in tensor factorization — tensor completion via certain type of non-linear neural network. We circumvent the notorious difficulty of tensor problems by adopting a dynamical systems perspective, and characterizing the evolution induced by gradient descent. The characterization suggests a form of greedy low tensor rank search, which we rigorously prove under certain conditions, and empirically demonstrate under others. Motivated by tensor rank capturing the implicit regularization of a non-linear neural network, we empirically explore it as a measure of complexity, and find that it captures the essence of datasets on which neural networks generalize. This leads us to believe that tensor rank may pave way to explaining both implicit regularization in deep learning, and the properties of real-world data translating this implicit regularization to generalization.

1 Introduction

The ability of neural networks to generalize when having far more learnable parameters than training examples, even in the absence of any explicit regularization, is an enigma lying at the heart of deep learning theory. Conventional wisdom is that this generalization stems from an implicit regularization — a tendency of gradient-based optimization to fit training examples with predictors whose “complexity” is as low as possible. The fact that “natural” data gives rise to generalization while other types of data (e.g. random) do not, is understood to result from the former being amenable to fitting by predictors of lower complexity. A major challenge in formalizing this intuition is that we lack definitions for predictor complexity that are both quantitative (i.e. admit quantitative generalization bounds) and capture the essence of natural data (in the sense of it being fittable with low complexity). Consequently, existing analyses typically focus on simplistic settings, where a notion of complexity is apparent. A prominent example of such a setting is matrix completion.

In matrix completion, we are given a randomly chosen subset of entries from an unknown matrix $W^* \in \mathbb{R}^{d,d}$, and our goal is to recover unseen entries. This can be viewed as a prediction problem, where the set of possible inputs is $\mathcal{X} = \{1, \ldots, d\} \times \{1, \ldots, d'\}$, the possible labels are $\mathcal{Y} = \mathbb{R}$, and the label of $(i, j) \in \mathcal{X}$ is $[W^*]_{i,j}$. Under this viewpoint, observed entries constitute the training set, and the average reconstruction error over unobserved entries is the test error, quantifying generalization. A predictor, i.e. a function from $\mathcal{X}$ to $\mathcal{Y}$, can then be seen as a matrix, and a natural notion of complexity is its rank. It is known empirically (cf. Gunasekar et al. (2017); Arora et al. (2019)) that this complexity measure is oftentimes implicitly minimized by matrix factorization — linear neural network trained via gradient descent with small learning rate and near-zero initialization. Mathematically characterizing the implicit regularization in matrix factorization is a highly active area of research. Though initially conjectured to be equivalent to norm minimization (see Gunasekar et al. (2017)), recent studies (Arora et al., 2019; Razin & Cohen, 2020; Li et al., 2021) suggest that this is not the case, and instead adopt a dynamical view, ultimately establishing that (under certain conditions) the implicit regularization in matrix factorization is performing a greedy low rank search.

A central question that arises is the extent to which the study of implicit regularization in matrix factorization is relevant to more practical settings. Recent experiments (see Razin & Cohen (2020)) have shown that the tendency towards low rank extends from matrices (two-dimensional arrays) to tensors (multi-dimensional arrays). Namely, in the task of $N$-dimensional tensor completion, which (analogously

---

1Equal contribution 1 Blavatnik School of Computer Science, Tel Aviv University, Israel. Correspondence to: Noam Razin <noam.razin@cs.tau.ac.il>, Asaf Maman <asafmaman@mail.tau.ac.il>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).
to matrix completion) can be viewed as a prediction problem over $N$ input variables, training a tensor factorization via gradient descent with small learning rate and near-zero initialization tends to produce tensors (predictors) with low tensor rank. Analogously to how matrix factorization may be viewed as a linear neural network, tensor factorization can be seen as a certain type of non-linear neural network (two-layer network with multiplicative non-linearity, cf. Cohen et al. (2016b)), and so it represents a setting much closer to practical deep learning.

In this paper we provide the first theoretical analysis of implicit regularization in tensor factorization. We circumvent the notorious difficulty of tensor problems (see Hillar & Lim (2013)) by adopting a dynamical systems perspective. Characterizing the evolution that gradient descent with small learning rate and near-zero initialization induces on the components of a factorization, we show that their norms are subject to a momentum-like effect, in the sense that they move slower when small and faster when large. This implies a form of greedy low tensor rank search, generalizing phenomena known for the case of matrices. We employ the finding to prove that, with the classic Huber loss from robust statistics (Huber, 1964), arbitrarily small initialization leads tensor factorization to follow a trajectory of rank one tensors for an arbitrary amount of time or distance. Experiments validate our analysis, demonstrating implicit regularization towards low tensor rank in a wide array of configurations.

Motivated by the fact that tensor rank captures the implicit regularization of a non-linear neural network, we empirically explore its potential to serve as a measure of complexity for multivariable predictors. We find that it is possible to fit standard image recognition datasets — MNIST (Le-Cun, 1998) and Fashion-MNIST (Xiao et al., 2017) — with predictors of extremely low tensor rank, far beneath what is required for fitting random data. This leads us to believe that tensor rank (or more advanced notions such as hierarchical tensor ranks) may pave way to explaining both implicit regularization of contemporary deep neural networks, and the properties of real-world data translating this implicit regularization to generalization.

The remainder of the paper is organized as follows. Section 2 presents the tensor factorization model, as well as its interpretation as a neural network. Section 3 characterizes its dynamics, followed by Section 4 which employs the characterization to establish (under certain conditions) implicit tensor rank minimization. Experiments, demonstrating both the dynamics of learning and the ability of tensor rank to capture the essence of standard datasets, are given in Section 5. In Section 6 we review related work. Finally, Section 7 concludes. Extension of our results to tensor sensing (more general setting than tensor completion) is discussed in Appendix A.

## 2 Tensor Factorization

Consider the task of completing an $N$-dimensional tensor ($N \geq 3$) with axis lengths $d_1, \ldots, d_N \in \mathbb{N}$, or, in standard tensor analysis terminology, an order $N$ tensor with modes of dimensions $d_1, \ldots, d_N$. Given a set of observations \( \{y_{i_1, \ldots, i_N} \in \mathbb{R}\}_{(i_1, \ldots, i_N) \in \Omega} \), where $\Omega$ is a subset of all possible index tuples, a standard (undetermined) loss function for the task is:

$$
\mathcal{L} : \mathbb{R}^{d_1, \ldots, d_N} \rightarrow \mathbb{R}_{\geq 0}
$$

(1)

$$
\mathcal{L}(\mathcal{W}) = \frac{1}{|\Omega|} \sum_{(i_1, \ldots, i_N) \in \Omega} \ell(\{y_{i_1, \ldots, i_N} - y_{i_1, \ldots, i_N}\}),
$$

where $\ell : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}$ is differentiable and locally smooth. A typical choice for $\ell(\cdot)$ is $\ell(z) = \frac{1}{2}z^2$, corresponding to $\ell_2$ loss. Other options are also common, for example that given in Equation (8), which corresponds to the Huber loss from robust statistics (Huber, 1964) — a differentiable surrogate for $\ell_1$ loss.

Performing tensor completion with an $R$-component tensor factorization amounts to optimizing the following (non-convex) objective:

$$
\phi(\{w^R_{n}\}_{r=1}^{N}) := \mathcal{L}(\mathcal{W}_e),
$$

(2)

defined over weight vectors \( \{w^R_{n}\}_{r=1}^{N} \subseteq \mathbb{R}^{d_1, \ldots, d_N} \), where:

$$
\mathcal{W}_e := \sum_{r=1}^{R} w^r_{1} \otimes \cdots \otimes w^r_{N}
$$

(3)

is referred to as the end tensor of the factorization, with $\otimes$ representing outer product. The minimal number of components $R$ required in order for $\mathcal{W}_e$ to be able to express a given tensor $\mathcal{V} \in \mathbb{R}^{d_1, \ldots, d_N}$, is defined to be the tensor rank of $\mathcal{V}$. One may explicitly restrict the tensor rank of solutions produced by the tensor factorization via limiting $R$. However, since our interest lies in the implicit regularization induced by gradient descent, i.e. in the type of end tensors (Equation (3)) it will find when applied to the objective $\phi(\cdot)$ (Equation (2)) with no explicit constraints, we treat the case where $R$ can be arbitrarily large.

In line with analyses of matrix factorization (e.g. Gunasekar et al. (2017); Arora et al. (2018; 2019); Eftekhari & Zygalakis (2020); Li et al. (2021)), we model small learning rate for gradient descent through the infinitesimal limit,
Implicit Regularization in Tensor Factorization

Figure 1: Prediction tasks over discrete variables can be viewed as tensor completion problems. Consider the task of learning a predictor from domain \(X = \{1, \ldots, d_1\} \times \cdots \times \{1, \ldots, d_N\}\) to range \(Y = \mathbb{R}\) (figure assumes \(N = 3\) and \(d_1 = \cdots = d_N = 5\) for the sake of illustration). Each input sample is associated with a location in an order \(N\) tensor with mode (axis) dimensions \(d_1, \ldots, d_N\), where the value of a variable (depicted as a shade of gray) determines the index of the corresponding mode (marked by “A”, “B” or “C”). The associated location stores the label of the sample. Under this viewpoint, training samples are observed entries, drawn according to an unknown distribution from a ground truth tensor. Learning a predictor amounts to completing the unobserved entries, with test error measured by (weighted) average reconstruction error.

In many standard prediction tasks (e.g. image recognition), only a small subset of the input domain has non-negligible probability. From the tensor completion perspective this means that observed entries reside in a restricted part of the tensor, and reconstruction error is weighted accordingly (entries outside the support of the distribution are neglected).

### 2.1 Interpretation as Neural Network

Tensor completion can be viewed as a prediction problem, where each mode corresponds to a discrete input variable. For an unknown tensor \(W^* \in \mathbb{R}^{d_1 \times \cdots \times d_N}\), inputs are index tuples of the form \((i_1, \ldots, i_N)\), and the label associated with such an input is \([W^*]_{i_1,\ldots,i_N}\). Under this perspective, the training set consists of the observed entries, and the average reconstruction error over unseen entries measures test error. The standard case, in which observations are drawn uniformly across the tensor and reconstruction error weighs all entries equally, corresponds to a data distribution that is uniform, but other distributions are also viable.

Consider for example the task of predicting a continuous label for a 100-by-100 binary image. This can be formulated as an order 10000 tensor completion problem, where all modes are of dimension 2. Each input image corresponds to a location (entry) in the tensor \(W^*\), holding its continuous label. As image pixels are (typically) not distributed independently and uniformly, locations in the tensor are not drawn uniformly when observations are generated, and are not weighted equally when reconstruction error is computed. See Figure 1 for further illustration of how a general prediction task (with discrete inputs and scalar output) can be formulated as a tensor completion problem.

Under the above formulation, tensor factorization can be viewed as a two layer neural network with multiplicative non-linearity. Given an input, \(i.e.\) a location in the tensor, the network produces an output equal to the value that the factorization holds at the given location. Figure 2 illustrates this equivalence between solving tensor completion with a tensor factorization and solving a prediction problem with a non-linear neural network. A major drawback of matrix factorization as a theoretical surrogate for modern deep learning is that it misses the critical aspect of non-linearity. Tensor factorization goes beyond the realm of linear predictors — a significant step towards practical neural networks.

### 3 Dynamical Characterization

In this section we derive a dynamical characterization for the norms of individual components in the tensor factorization. The characterization implies that with small learning rate and near-zero initialization, components tend to be learned incrementally, giving rise to a bias towards low tensor rank solutions. This finding is used in Section 4 to prove (under certain conditions) implicit tensor rank minimization, and is demonstrated empirically in Section 5.

Hereafter, unless specified otherwise, when referring to a norm we mean the standard Frobenius (Euclidean) norm, denoted by \(\|\cdot\|\).

The following lemma establishes an invariant of the dynamics, showing that the differences between squared norms of vectors in the same component are constant through time.

**Lemma 1.** For all \(r \in \{1, \ldots, R\}\) and \(n, \bar{n} \in \{1, \ldots, N\}\):

\[
\|w^r_n(t)\|^2 - \|w^r_{\bar{n}}(t)\|^2 = \|w^r_n(0)\|^2 - \|w^r_{\bar{n}}(0)\|^2, \quad t \geq 0.
\]

**Proof sketch (for proof see Lemma 9 in Subappendix C.2.2).**

The claim readily follows by showing that under gradient flow \(\frac{d}{dt}\|w^r_n(t)\|^2 = \frac{d}{dt}\|w^r_{\bar{n}}(t)\|^2\) for all \(t \geq 0\). \(\square\)

\(\text{We note that all results in this section apply even if the tensor completion loss } L(\cdot) \text{ (Equation (1)) is replaced by any differentiable and locally smooth function. The proofs in Appendix C already account for this more general setting.}\)
Lemma 1 naturally leads to the definition below.

**Definition 1.** The unbalancedness magnitude of the weight vectors \( \{ w^r_n \in \mathbb{R}^{d_r} \}_{r=1}^R \) is defined to be:

\[
\max_{r \in \{1, \ldots, R\}, \ n, \tilde{n} \in \{1, \ldots, N\}} \left( \| w^r_n \|^2 - \| w^r_{\tilde{n}} \|^2 \right).
\]

By Lemma 1, the unbalancedness magnitude is constant during optimization, and thus, is determined at initialization. When weight vectors are initialized near the origin — regime of interest — the unbalancedness magnitude is small, approaching zero as initialization scale decreases.

Theorem 1 below provides a dynamical characterization for norms of individual components in the tensor factorization.

**Theorem 1.** Assume unbalancedness magnitude \( \varepsilon \geq 0 \) at initialization, and denote by \( W(t) \) the end tensor (Equation (3)) at time \( t \geq 0 \) of optimization. Then, for every \( r \in \{1, \ldots, R\} \) and time \( t \geq 0 \) at which \( \| \otimes_{n=1}^N w^r_n(t) \| > 0 \):

- If \( \gamma_r(t) := \langle -\nabla L(W(t)), \otimes_{n=1}^N w^r_n(t) \rangle \geq 0 \), then:

\[
\frac{d}{dt} \| \otimes_{n=1}^N w^r_n(t) \| \leq N \gamma_r(t) \left( \| \otimes_{n=1}^N w^r_n(t) \|^{2 \varepsilon} + \varepsilon \right)^{N-1}.
\]

- Otherwise, if \( \gamma_r(t) < 0 \), then:

\[
\frac{d}{dt} \| \otimes_{n=1}^N w^r_n(t) \| \geq N \gamma_r(t) \cdot \left( \frac{\| \otimes_{n=1}^N w^r_n(t) \|^{2 \varepsilon}}{\| \otimes_{n=1}^N w^r_n(t) \|^{2 \varepsilon} + \varepsilon} \right)^{N-1}.
\]

where \( \tilde{w}^r_n(t) := w^r_n(t) / \| w^r_n(t) \| \) for \( n = 1, \ldots, N \).

**Proof sketch (for proof see Subappendix C.3).** Differentiating a component’s norm with respect to time, we obtain

\[
\frac{d}{dt} \| \otimes_{n=1}^N w^r_n(t) \| = \gamma_r(t) \cdot \sum_{n'=1}^N \| w^r_{n'}(t) \|^2.
\]

The desired bounds then follow from using conservation of unbalancedness magnitude (as implied by Lemma 1), and showing that \( \| w^r_{n'}(t) \|^2 \leq \| \otimes_{n=1}^N w^r_n(t) \|^{2/N} + \varepsilon \) for all \( t \geq 0 \) and \( n' \in \{1, \ldots, N\} \).

Theorem 1 shows that when unbalancedness magnitude at initialization (denoted \( \varepsilon \)) is small, the evolution rates of component norms are roughly proportional to their size exponentiated by \( 2 - 2/N \), where \( N \) is the order of the tensor factorization. Consequently, component norms are subject to a momentum-like effect, by which they move slower when small and faster when large. This suggests that when initialized near zero, components tend to remain close to the origin, and then, upon reaching a critical threshold, quickly grow until convergence, creating an incremental learning effect that yields implicit regularization towards low tensor rank. This phenomenon is used in Section 4 to formally prove (under certain conditions) implicit tensor rank minimization, and is demonstrated empirically in Section 5.

When the unbalancedness magnitude at initialization is exactly zero, our dynamical characterization takes on a particularly lucid form.

**Corollary 1.** Assume unbalancedness magnitude zero at initialization. Then, with notations of Theorem 1, for any \( r \in \{1, \ldots, R\} \), the norm of the \( r \)th component evolves by:

\[
\frac{d}{dt} \| \otimes_{n=1}^N w^r_n(t) \| = \gamma_r(t) \cdot \| \otimes_{n=1}^N w^r_n(t) \|^{2 - 2 \varepsilon}.
\]

where by convention \( \tilde{w}^r_n(t) = 0 \) if \( w^r_n(t) = 0 \).
We will assume it is smaller than observed tensor entries:

Assumption 1.

where Arora et al. (2019) for singular values of a matrix factorization from incremental learning dynamics similar to above, with singular values standing in place of component norms. In Hillar & Lim (2013), cf. Huber loss from robust statistics (Huber, 1964), given by:

For technical reasons, our focus in this section lies on the Huber loss, in which case δ is predetermined. Huber loss is often used as a differentiable surrogate for the shifted one h loss, in which case

δ

where by definition

and let \{w^n_r(t)\}_{r,n} evolve per Equation (4). Denote by \{W_r(t), t \geq 0\}, the trajectory induced on the end tensor (Equation (3)). We will study the evolution of this trajectory through time. A hurdle that immediately arises is that, by the dynamical characterization of Section 3, when the initialization scale \( \alpha \) tends to zero (regime of interest), the time it takes \( W_r(t) \) to escape the origin grows to infinity. We overcome this hurdle by considering a reference sphere — a sphere around the origin with sufficiently small radius:

where \( \rho \) is chosen arbitrarily. With the reference sphere \( S \) at hand, we define a time-shifted version of the trajectory \( W_r(t) \), aligning \( t = 0 \) with the moment at which \( S \) is reached:

We will establish proximity of \( \bar{W}_r(t) \) to trajectories of rank one tensors. We say that \( W_r(t) \) is a rank one trajectory, if it coincides with some trajectory

Assumption 2. The initialization \( \{a^n_r\}_{r=1}^{R} \) has unbalancedness magnitude zero.

We further assume that within \( \{a^n_r\}_{r,n} \) there exists a leading component (subset \( \{a^n_{1\ldots R}\}_n \)), in the sense that it is larger than others, while having positive projection on the attracting force at the origin, i.e. on minus the gradient of the loss \( L(\cdot) \) (Equation (1)) at zero:

Assumption 3. There exists \( \bar{r} \in \{1, \ldots, R\} \) such that:

We will establish proximity of \( \bar{W}_r(t) \) to trajectories of rank one tensors starting from the initialization \( \{a^n_r\}_{r,n} \) scaled by \( \alpha \). That is, we set:

Note that this entails assumption of non-zero observations.

It is worthwhile highlighting the relation to matrix factorization. There, an implicit bias towards low rank emerges from incremental learning dynamics similar to above, with singular values standing in place of component norms. In fact, the dynamical characterization given in Corollary 1 is structurally identical to the one provided by Theorem 3 in Arora et al. (2019) for singular values of a matrix factorization. We thus obtained a generalization from matrices to tensors, notwithstanding the notorious difficulty often associated with the latter (cf. Hillar & Lim (2013)).

4 Implicit Tensor Rank Minimization

In this section we employ the dynamical characterization derived in Section 3 to theoretically establish implicit regularization towards low tensor rank. Specifically, we prove that under certain technical conditions, arbitrarily small initialization leads tensor factorization to follow a trajectory of rank one tensors for an arbitrary amount of time or distance. As a corollary, we obtain that if the tensor completion problem admits a rank one solution, and all rank one trajectories uniformly converge to it, tensor factorization with infinitesimal initialization will converge to it as well. Our analysis generalizes to tensor factorization recent results developed in Li et al. (2021) for matrix factorization. As typical in transitioning from matrices to tensors, this generalization entails significant challenges necessitating use of fundamentally different techniques.

For technical reasons, our focus in this section lies on the Huber loss from robust statistics (Huber, 1964), given by:

where \( \delta_h > 0 \), referred to as the transition point of the loss, is predetermined. Huber loss is often used as a differentiable surrogate for \( \ell_1 \) loss, in which case \( \delta_h \) is chosen to be small. We will assume it is smaller than observed tensor entries:

Assumption 1. \( \delta_h < |y_{i_1 \ldots i_N}|, \forall (i_1 \ldots i_N) \in \Omega \).

We will consider an initialization \( \{a^n_r \in \mathbb{R}^{d_1 \ldots d_N}\}_{r=1}^{R} \) for the weight vectors of the tensor factorization, and will scale this initialization towards zero. In line with infinitesimal initializations being captured by unbalancedness magnitude zero (cf. Section 3), we assume that this is the case:

Proof sketch (for proof see Subappendix C.4). If the time \( t \) is such that \( \|\otimes_{n=1}^{N} w^n_r(t)\| > 0 \), Equation (7) readily follows from applying Theorem 1 with \( \epsilon = 0 \). For the case where \( \|\otimes_{n=1}^{N} w^n_r(t)\| = 0 \), we show that the component \( \otimes_{n=1}^{N} w^n_r(t) \) must be identically zero throughout, hence both sides of Equation (7) are equal to zero.

\[ \ell_h : \mathbb{R} \rightarrow \mathbb{R}_{\geq 0}, \ell_h(z) := \begin{cases} \frac{1}{2}z^2, & |z| < \delta_h \\ 2\delta_h(|z| - \frac{1}{2}\delta_h), & \text{otherwise} \end{cases}, \]
We are now in a position to state our main result, by which arbitrarily small initialization leads tensor factorization to follow a (balanced) rank one trajectory for an arbitrary amount of time or distance.

**Theorem 2.** Under Assumptions 1, 2 and 3, for any distance from origin \( D > 0 \), time duration \( T > 0 \), and degree of approximation \( \epsilon \in (0, 1) \), if initialization scale \( \alpha \) is sufficiently small,\(^9\) then: (i) \( W_e(t) \) reaches the reference sphere \( S \); and (ii) there exists a balanced rank one trajectory \( W_1(t) \) emanating from \( S \), such that \( \| W_e(t) - W_1(t) \| \leq \epsilon \) at least until \( t \geq T \) or \( \| W_e(t) \| \geq D \).

Proof sketch (for proof see Subappendix C.5). Using the dynamical characterization from Section 3 (Lemma 1 and Corollary 1), and the fact that \( \nabla L(\cdot) \) is locally constant around the origin, we establish that (i) \( W_e(t) \) reaches the reference sphere \( S \); and (ii) at that time, the norm of the \( r \)th component is of constant scale (independent of \( \alpha \)), while the norms of all other components are \( \mathcal{O}(\alpha^N) \). Thus, taking \( \alpha \) towards zero leads \( W_e(t) \) to arrive at \( S \) while being arbitrarily close to the initialization of a balanced rank one trajectory \( W_1(t) \). Since the objective is locally smooth, this ensures \( W_e(t) \) is within distance \( \epsilon \) from \( W_1(t) \) for an arbitrary amount of time or distance. That is, if \( \alpha \) is sufficiently small, \( \| W_e(t) - W_1(t) \| \leq \epsilon \) at least until \( t \geq T \) or \( \| W_e(t) \| \geq D \).

As an immediate corollary of Theorem 2, we obtain that if all balanced rank one trajectories uniformly converge to a global minimum, tensor factorization with infinitesimal initialization will do so too. In particular, its implicit regularization will direct it towards a solution with tensor rank one.

**Corollary 2.** Assume the conditions of Theorem 2 (Assumptions 1, 2 and 3), and in addition, that all balanced rank one trajectories emanating from \( S \) converge to a tensor \( W^* \in \mathbb{R}^{d_1 \cdots d_N} \) uniformly, in the sense that they are all confined to some bounded domain, and for any \( \epsilon > 0 \), there exists a time \( T \) after which they are all within distance \( \epsilon \) from \( W_* \). Then, for any \( \epsilon > 0 \), if initialization scale \( \alpha \) is sufficiently small, there exists a time \( T \) for which \( \| W_e(T) - W_* \| \leq \epsilon \).

\(^8\)Note that the definitions of rank one trajectory and balanced rank one trajectory allow for \( W_1(t) \) to have rank zero (i.e. to be equal to zero) at some or all times \( t \geq 0 \).

\(^9\)Hiding problem-dependent constants, an initialization scale of \( \epsilon D^{-1} \exp(-\mathcal{O}(D^2 T)) \) suffices. Exact constants are specified at the beginning of the proof in Subappendix C.5.

5 Experiments

In this section we present our experiments. Subsection 5.1 corroborates our theoretical analyses (Sections 3 and 4), evaluating tensor factorization (Section 2) on synthetic low (tensor) rank tensor completion problems. Subsection 5.2 explores tensor rank as a measure of complexity, examining its ability to capture the essence of standard datasets. For brevity, we defer a description of implementation details, as well as some experiments, to Appendix B.

5.1 Dynamics of Learning

Recently, Razin & Cohen (2020) empirically showed that, with small learning rate and near-zero initialization, gradient descent over tensor factorization exhibits an implicit regularization towards low tensor rank. Our theory (Sections 3 and 4) explains this implicit regularization through a dynamical analysis — we prove that the movement of component norms is attenuated when small and enhanced when large, thus creating an incremental learning effect which becomes more potent as initialization scale decreases. Figure 3 demonstrates this phenomenon empirically on synthetic low (tensor) rank tensor completion problems. Figures 5, 6 and 7 in Subappendix B.1 extend the experiment, corroborating our analyses in a wide array of settings.

5.2 Tensor Rank as Measure of Complexity

Implicit regularization in deep learning is typically viewed as a tendency of gradient-based optimization to fit training examples with predictors whose “complexity” is as low as possible. The fact that “natural” data gives rise to generalization while other types of data (e.g. random) do not, is understood to result from the former being amenable to fitting by predictors of lower complexity. A major challenge in formalizing this intuition is that we lack definitions for predictor complexity that are both quantitative (i.e. admit quantitative generalization bounds) and capture the essence of natural data (types of data on which neural networks generalize in practice), in the sense of it being fittable with low complexity.

As discussed in Subsection 2.1, learning a predictor with multiple discrete input variables and a continuous output can be viewed as a tensor completion problem. Specifically,
As representatives of natural data, we chose the classic
with

\[ n \in \mathbb{N}, d_1, \ldots, d_N \in \mathbb{N} \]

learning a predictor from domain
\[ X = \{1, \ldots, d_1\} \times \cdots \times \{1, \ldots, d_N\} \]
to range \[ Y = \mathbb{R} \]
corresponds to completion of an order \( N \) tensor with mode (axis) dimensions \( d_1, \ldots, d_N \). Under this correspondence, any predictor can simply be thought of as a tensor, and vice versa. We have shown that solving tensor completion via tensor factorization amounts to learning a predictor through a certain neural network (Subsection 2.1), whose implicit regularization favors solutions with low tensor rank (Sections 3 and 4). Motivated by these connections, the current subsection empirically explores tensor rank as a measure of complexity for predictors, by evaluating the extent to which it captures natural data, i.e. allows the latter to be fit with low complexity predictors.

As representatives of natural data, we chose the classic MNIST dataset (LeCun, 1998) — perhaps the most common benchmark for demonstrating ideas in deep learning — and its more modern counterpart Fashion-MNIST (Xiao et al., 2017). A hurdle posed by these datasets is that they involve classification into multiple categories, whereas the equivalence to tensors applies to predictors whose output is a scalar. It is possible to extend the equivalence by equating a multi-output predictor with multiple tensors, in which case the predictor is associated with multiple tensor ranks. However, to facilitate a simple presentation, we avoid this extension and simply map each dataset into multiple one-vs-all binary classification problems. For each problem, we associate the label 1 with the active category and 0 with all the rest, and then attempt to fit training examples with predictors of low tensor rank, reporting the resulting mean squared error, i.e. the residual of the fit. This is compared against residuals obtained when fitting two types of random data: one generated via shuffling labels, and the other by replacing inputs with noise.

Both MNIST and Fashion-MNIST comprise 28-by-28 grayscale images, with each pixel taking one of 256 possible values. Tensors associated with predictors are thus of order 784, with dimension 256 in each mode (axis).\(^a\) A general rank one tensor can then be expressed as an outer product between 784 vectors of dimension 256 each, and accordingly has roughly \( 784 \cdot 256 \) degrees of freedom. This significantly exceeds the number of training examples in the datasets (60000), hence it is no surprise that we could easily fit them, as well as their random variants, with a predictor whose tensor rank is one. To account for the comparatively small training sets, and render their fit more challenging, we quantized pixels to hold one of two values, i.e. we reduced images from grayscale to black and white. Following the quantization, tensors associated with predictors have dimension two in each mode, and the number of degrees of freedom in a general rank one tensor is roughly \( 784 \cdot 2 \) — well below the number of training examples. We may thus expect to see a difference between the tensor ranks needed for fitting original datasets and those required by the random

\(^a\)In practice, when associating predictors with tensors, it is often beneficial to modify the representation of the input (cf. Cohen et al. (2016b)). For example, in the context under discussion, rather than having the discrete input variables hold pixel intensities, they may correspond to small image patches, where each patch is represented by the index of a centroid it is closest to, with centroids determined via clustering applied to all patches across all images in the dataset. For simplicity, we did not transform representations in our experiments, and simply operated over raw image pixels.

Figure 3: Dynamics of gradient descent over tensor factorization — incremental learning of components yields low tensor rank solutions. Presented plots correspond to the task of completing a (tensor) rank 5 ground truth tensor of size 10-by-10-by-10-by-10 (order 4) based on 2000 observed entries chosen uniformly at random without repetition (smaller sample sizes led to solutions with tensor rank lower than that of the ground truth tensor). In each experiment, the \( \ell_2 \) loss (more precisely, Equation (1) with \( \ell(z) = z^2 \)) was minimized via gradient descent over a tensor factorization with \( R = 1000 \) components (large enough to express any tensor), starting from (small) random initialization. First (left) three plots show (Frobenius) norms of the ten largest components under three standard deviations for initialization — 0.05, 0.01, and 0.005. Further reduction of initialization scale yielded no noticeable change. The rightmost plot compares reconstruction errors (Frobenius distance from ground truth) from the three runs. To facilitate more efficient experimentation, we employed an adaptive learning rate scheme (see Subappendix B.2 for details). Notice that, in accordance with the theoretical analysis of Section 3, component norms move slower when small and faster when large, creating an incremental process in which components are learned one after the other. This effect is enhanced as initialization scale is decreased, producing low tensor rank solutions that accurately reconstruct the low (tensor) rank ground truth tensor. In particular, even though the factorization consists of 1000 components, when initialization is sufficiently small, only five (tensor rank of the ground truth tensor) substantially depart from zero. Appendix B provides further implementation details, as well as similar experiments with: (i) Huber loss (see Equation (8)) instead of \( \ell_2 \) loss; (ii) ground truth tensors of different orders and (tensor) ranks; and (iii) tensor sensing (see Appendix A).
Theoretical analysis of implicit regularization induced by certain conditions, implicit minimization of a complexity measure or approximately.\footnote{Recent results of Vardi & Shamir (2021) imply that under certain conditions, implicit minimization of a complexity measure must be approximate (cannot be exact).} The latter type of results is perhaps the most common, covering complexity measures based on: frequency content of input-output mapping (Rahaman et al., 2019; Xu, 2018); curvature of training objective (Mudayoff & Michaeli, 2020); and norm or margin of weights or input-output mapping (Soudry et al., 2018; Gunasekar et al., 2018a;b; Jacot et al., 2018; Ji & Telgarsky, 2019b; Mei et al., 2019; Wu et al., 2020; Nacson et al., 2019; Ji & Telgarsky, 2019a; Oymak & Soltanolkotabi, 2019; Ali et al., 2020; Woodworth et al., 2020; Chizat & Bach, 2020; Yun et al., 2021). An additional complexity measure, arguably the most extensively studied, is matrix rank.

Rank minimization in matrix completion (or sensing) is a classic problem in science and engineering (cf. Davenport & Romberg (2016)). It relates to deep learning when solved via linear neural network, \textit{i.e.} through matrix factorization. The literature on matrix factorization for rank minimization is far too broad to cover here --- we refer to Chi et al. (2019) for a recent review. Notable works proving rank minimization via matrix factorization trained by gradient descent with no explicit regularization are Tu et al. (2016); Ma et al. (2018); Li et al. (2018). Gunasekar et al. (2017) conjectured that this implicit regularization is equivalent to norm minimization, but the recent studies Arora et al. (2019); Razin & Cohen (2020); Li et al. (2021) argue otherwise, and instead adopt a dynamical view, ultimately establishing that (under certain conditions) the implicit regularization in matrix factorization is performing a greedy low rank search. These studies are relevant to ours in the sense that we generalize some of their results to tensor factorization. As typical in transitioning from matrices to tensors (see Hillar & Lim & Romberg (2016)).
We note that the equivalence between tensor factorizations and different types of neural networks has been studied extensively, primarily in the context of expressive power (see, e.g., Cohen et al. (2016b); Cohen & Shashua (2016); Sharir et al. (2016); Cohen & Shashua (2017); Cohen et al. (2017); Sharir & Shashua (2018); Levine et al. (2018b); Cohen et al. (2018); Levine et al. (2018a); Balda et al. (2018); Khrulkov et al. (2018); Levine et al. (2019); Khrulkov et al. (2019); Levine et al. (2020)). Connections between tensor analysis and generalization in deep learning have also been made (cf. Li et al. (2020)), but to the best of our knowledge, the notion of quantifying the complexity of predictors through their tensor rank (supported empirically in Subsection 5.2) is novel to this work.

7 Conclusion

In this paper we provided the first theoretical analysis of implicit regularization in tensor factorization. To circumvent the notorious difficulty of tensor problems (see Hillar & Lim (2013)), we adopted a dynamical systems perspective, and characterized the evolution that gradient descent (with small learning rate and near-zero initialization) induces on the components of a factorization. The characterization suggests a form of greedy low tensor rank search, rigorously proven under certain conditions. Experiments demonstrated said phenomena.

A major challenge in mathematically explaining generalization in deep learning is to define measures for predictor complexity that are both quantitative (i.e. admit quantitative generalization bounds) and capture the essence of “natural” data (types of data on which neural networks generalize in practice), in the sense of it being fittable with low complexity. Motivated by the fact that tensor factorization is equivalent to a certain non-linear neural network, and by our analysis implying that the implicit regularization of this network minimizes tensor rank, we empirically explored the potential of the latter to serve as a measure of predictor complexity. We found that it is possible to fit standard image recognition datasets (MNIST and Fashion-MNIST) with predictors of extremely low tensor rank (far beneath what is required for fitting random data), suggesting that it indeed captures aspects of natural data.

The neural network to which tensor factorization is equivalent entails multiplicative non-linearity. It was shown in Cohen & Shashua (2016) that more prevalent non-linearities, for example rectified linear unit (ReLU), can be accounted for by considering generalized tensor factorizations. Studying the implicit regularization in generalized tensor factorizations (both empirically and theoretically) is regarded as a promising direction for future work.

There are two drawbacks to tensor factorization when applied to high-dimensional prediction problems. The first is technical, and relates to numerical stability — an order $N$ tensor factorization involves products of $N$ numbers, thus is susceptible to arithmetic underflow or overflow if $N$ is large. Care should be taken to avoid this pitfall, for example by performing computations in log domain (as done in Cohen & Shashua (2014); Cohen et al. (2016a); Sharir et al. (2016)). The second limitation is more fundamental, arising from the fact that tensor rank — the complexity measure implicitly minimized — is oblivious to the ordering of tensor modes (axes). This means that the implicit regularization does not take into account how predictor inputs are arranged (e.g., in the context of image recognition, it does not take into account spatial relationships between pixels). A potentially promising path for overcoming this limitation is introduction of hierarchy into the tensor factorization, equivalent to adding depth to the corresponding neural network (cf. Cohen et al. (2016b)). It may then be the case that a hierarchical tensor rank (see Hackbusch (2012)), which does account for mode ordering, will be implicitly minimized. We hypothesize that hierarchical tensor ranks may be key to explaining both implicit regularization of contemporary deep neural networks, and the properties of real-world data translating this implicit regularization to generalization.

Acknowledgements

This work was supported by a Google Research Scholar Award, a Google Research Gift, the Yandex Initiative in Machine Learning, Len Blavatnik and the Blavatnik Family Foundation, and Amnon and Anat Shashua.
References


Implicit Regularization in Tensor Factorization


Implicit Regularization in Tensor Factorization


Wang, X., Wu, C., Lee, J. D., Ma, T., and Ge, R. Beyond lazy training for over-parameterized tensor decomposition. 2020.


