Implicit regularization of gradient descent

Jingxing Wang and Mriganka Basu Roy Chowdhury

May 2024

Contents

1 Gradient descent and gradient flow

Given a sufficiently smooth function $f : \mathbb{R}^d \to \mathbb{R}$, a common technique to minimize it (or equivalently, maximize it), is to use the first-order optimality condition

$$\nabla f(x) = 0.$$

(Strictly speaking, this is satisfied for every *critical point*). An algorithmic approach to using this fact is the all-important *gradient descent*, which is an iterative algorithm that starts with an estimate $x_0 \in \mathbb{R}^d$, and successively attempts to improve it in steps:

$$x_{n+1} = x_n - \eta \nabla f(x_n). \tag{1}$$

The tunable parameter η is called the *learning rate*. The hope is that this converges, as $n \to \infty$ to a good solution. Indeed, if it converges, the limit x^* must satisfy $\nabla f(x^*) = 0$, by taking limits on both sides of (1) (and assuming f' is continuous).

Due to the discrete nature of this algorithm, many attempts to study it proceeds via first smoothening the algorithm by sending $\eta \to 0$, producing what is called the *gradient flow*:

$$\dot{x}_t = -\nabla f(x_t), \tag{2}$$

an *autonomous* (hence the name *flow*) ordinary differential equation in $\mathbb{R}^d$. Although a significant section of the optimization literature is dedicated to advice on choosing the rate η and infinitesimal choices are not always optimal (for example, due to performance considerations), gradient flow captures a large portion of the overall behavior of gradient descent, and hence is extremely valuable to study.

2 Gradient flow in least squares

Recall the ordinary least squares problem, which arises (among other places) in the study of the multivariate linear regression problem:

$$\underset{\beta}{\operatorname{argmin}}\, f(\beta) = \underset{\beta}{\operatorname{argmin}}\, \|y - X\beta\|_2^2 \tag{3}$$

where $y \in \mathbb{R}^n, \beta \in \mathbb{R}^d$ and the design matrix $X \in \mathbb{R}^{n\times d}$. Rewriting the objective f in (3), we have

$$\begin{aligned} f(\beta) &= (y - X\beta)^\mathsf{T}(y - X\beta) \\ &= y^\mathsf{T} y - 2y^\mathsf{T} X\beta + \beta^\mathsf{T}(X^\mathsf{T} X)\beta \end{aligned}$$

Thus, $\nabla f(\beta)$ is

$$\begin{aligned} \nabla(-2y^\mathsf{T} X\beta + \beta^T(X^\mathsf{T} X)\beta) &\overset{(\alpha)}{=} -2(y^\mathsf{T} X)^\mathsf{T} + 2(X^\mathsf{T} X)\beta \\ &= -2\left[X^\mathsf{T} y - X^\mathsf{T} X\beta\right] \end{aligned} \tag{4}$$

where in step (α) we use the two facts

- $\nabla_x(a^\mathsf{T} x) = a$, where a is a vector, and
- $\nabla_x(x^\mathsf{T} Ax) = 2Ax$ if A is symmetric.

The gradient flow ODE in this case is given by (2) specialized to f to obtain

$$\dot{\beta}_t = -\nabla f(x_t) = 2\left[X^\mathsf{T} y - X^\mathsf{T} X\beta_t\right].$$

To reduce the notation a bit, let us set $2X^\mathsf{T} y = z$, and $2X^\mathsf{T} X = A$, so that the equation above is

$$\dot{\beta}_t = z - A\beta_t \tag{5}$$

To solve this equation in closed form, we will require the *matrix exponential*, defined for any square matrix $A \in \mathbb{R}^{d\times d}$ as

$$e^A \overset{\text{def}}{=} \sum_{k=0}^{\infty} \frac{A^k}{k!}$$

This converges for every A, as can be checked simply by using $\left\|A^k\right\|_{\text{HS}} \leq \|A\|_{\text{HS}}^k$ and the triangle inequality for $\|\cdot\|_{\text{HS}}$.

A key property of the exponential is that

$$\frac{d(e^{tA})}{dt} = Ae^{tA} = e^{tA}A.$$

Define $\theta_t = e^{tA}\beta_t$. Using the fact that for matrix valued functions M_t and vector valued function x_t, we have $(\dot{Mx})_t = \dot{M}_t x_t + M_t \dot{x}_t$, we see that

$$\begin{aligned}
\dot{\theta}_t = (e^{t\dot{A}}\beta_t) &= e^{tA}A\beta_t + e^{tA}\dot{\beta}_t \\
&= e^{tA}\left[\dot{\beta}_t + A\beta_t\right] \\
&\overset{\text{using (5)}}{=} e^{tA}z.
\end{aligned}$$

Thus by integrating,

$$\theta_t = \left(\int_0^t e^{sA}ds\right)z,$$

so that the overall solution is

$$\beta_t = e^{-tA}\left(\int_0^t e^{sA}ds\right)z \tag{6}$$

Suppose the SVD of X is $U\Sigma V^T$, where $U \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{d\times d}$ are orthogonal and $\Sigma \in \mathbb{R}^{n\times d}$ is diagonal with the singular values $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$ where $r = \operatorname{rank} X \leq \min(n,d)$. Let $\Lambda = \Sigma^T\Sigma$, and let define $\lambda_i \overset{\text{def}}{=} \Lambda_{ii} = \sigma_i^2$ for $i = 1,\ldots,d$. Observe that λ_i are nonincreasing and $\lambda_j = 0$ for all $j > r$.

Then

$$A = 2X^{\mathsf{T}}X = 2V\Sigma^TU^TU\Sigma V^T = 2V(\Sigma^T\Sigma)V^T,$$

and therefore,

$$e^{tA} = V\cdot\operatorname{diag}(e^{2t\lambda_1},\ldots,e^{2t\lambda_d})\cdot V^T$$

using the following fact about the matrix exponential: if $M = QDQ^T$ with Q orthogonal and D diagonal, then

$$e^{tM} = e^{tQDQ^T} = Qe^{tD}Q^T = Q\cdot\operatorname{diag}(e^{tD_{11}},\ldots,e^{tD_{dd}})\cdot Q^T.$$

Substituting these expressions in (6), we get

$$\begin{aligned}
\beta_t &= V\cdot\begin{bmatrix} e^{-2t\lambda_1} & & \\ & \ddots & \\ & & e^{-2t\lambda_d}\end{bmatrix}\cdot V^{\mathsf{T}}\cdot\left(\int_0^t V\cdot\begin{bmatrix} e^{2s\lambda_1} & & \\ & \ddots & \\ & & e^{2s\lambda_d}\end{bmatrix}\cdot V^{\mathsf{T}}ds\right)\cdot 2V\Sigma^{\mathsf{T}}U^{\mathsf{T}}y \\
&= V\cdot\begin{bmatrix} e^{-2t\lambda_1} & & \\ & \ddots & \\ & & e^{-2t\lambda_d}\end{bmatrix}\cdot\left(\int_0^t\begin{bmatrix} e^{2s\lambda_1} & & \\ & \ddots & \\ & & e^{2s\lambda_d}\end{bmatrix}ds\right)\cdot 2\Sigma^{\mathsf{T}}U^{\mathsf{T}}y
\end{aligned}$$

Now recall that $\lambda_{r+1} = \ldots = \lambda_d = 0$, and the rest are nonzero. So we may perform the integration to obtain

$$\begin{aligned}
&= 2V\cdot\left[\begin{array}{ccc|c} e^{-2t\lambda_1} & & & \\ & \ddots & & \\ & & e^{-2t\lambda_r} & \\ \hline & & & I_{d-r}\end{array}\right]\cdot\left[\begin{array}{ccc|c} \frac{e^{2t\lambda_1}-1}{2\lambda_1} & & & \\ & \ddots & & \\ & & \frac{e^{2t\lambda_r}-1}{2\lambda_r} & \\ \hline & & & t\cdot I_{d-r}\end{array}\right]\cdot\Sigma^{\mathsf{T}}U^{\mathsf{T}}y \\
&= 2V\cdot\left[\begin{array}{ccc|c} \frac{e^{2t\lambda_1}-1}{2\lambda_1e^{2t\lambda_1}} & & & \\ & \ddots & & \\ & & \frac{e^{2t\lambda_r}-1}{2\lambda_re^{2t\lambda_r}} & \\ \hline & & & t\cdot I_{d-r}\end{array}\right]\cdot\Sigma^{\mathsf{T}}U^{\mathsf{T}}y
\end{aligned}$$

Finally, recalling that Σ^{T} is a diagonal (but rectangular!) matrix of size $n \times d$, with σ_i on the diagonal, and $\sigma_i = 0$ if $i > r$, we obtain

$$= V \cdot \Sigma_t^{\mathsf{T}} \cdot U^{\mathsf{T}} y,$$

where $\Sigma_t \in \mathbb{R}^{n\times d}$ is the same shape as Σ but with the following diagonal elements

$$\begin{aligned}
&(\Sigma_t)_{ii} = \frac{e^{2t\lambda_i} - 1}{\lambda_i \cdot e^{2t\lambda_i}} \sigma_i = \frac{e^{2t\lambda_i} - 1}{e^{2t\lambda_i}} \cdot \lambda_i^{-1/2}, \quad i = 1, \dots, r \\
&(\Sigma_t)_{ii} = 0, \quad \text{elsewhere.}
\end{aligned}$$

In summary,

$$\beta_t = V \cdot \Sigma_t^{\mathsf{T}} \cdot U^{\mathsf{T}} y,$$

where Σ_t is defined above using the SVD $X = U\Sigma V^{\mathsf{T}}$ for the design matrix. This explicit solution will feature in our analysis below.

3 Implicit regularization in gradient flow

Observe that as $t \to \infty$, we have $\beta_t \to V \cdot (\Sigma^+)^{\mathsf{T}} \cdot U^{\mathsf{T}} y$, where

$$\begin{aligned}
&(\Sigma^+)_{ii} = \lambda_i^{-1/2} = 1/\sigma_i, \quad i = 1, \dots, r, \\
&(\Sigma^+)_{ii} = 0 \quad \text{otherwise.}
\end{aligned}$$

To interpret this behavior, let us return to the first order conditions defined in (4), according to which, a minimizer of the least squares objective must satisfy

$$X^{\mathsf{T}} X \beta = X^{\mathsf{T}} y.$$

Using the SVD of X, this is the same as

$$\begin{aligned}
&V\Sigma^T U^{\mathsf{T}} U \Sigma V^{\mathsf{T}} \beta = V \Sigma^{\mathsf{T}} U^{\mathsf{T}} y \\
\iff &V \Sigma^{\mathsf{T}} \Sigma V^{\mathsf{T}} \beta = V \Sigma^{\mathsf{T}} U^{\mathsf{T}} y \\
\iff &\Lambda V^{\mathsf{T}} \beta = \Sigma^{\mathsf{T}} U^{\mathsf{T}} y
\end{aligned}$$

Although $V^{\mathsf{T}} = V^{-1}$ is full rank, Λ is only of rank r which is potentially less than d. Thus, this equation can have more than one solution, since any vector θ satisfying $\Lambda V^{\mathsf{T}} \theta = 0$ produces another solution $\beta + \theta$ to the equation above.

In such a situation, one may wish to find a solution β which, *additionally, also has the least L^2 norm*. This is a form of *regularization*. Due to the strict convexity of the L^2 norm, and the linearity of the first order constraint above, this minimal L^2 solution will be unique. We will try to determine this solution now.

Observe that $\|\beta\|_2^2 = \left\|V^{\mathsf{T}}\beta\right\|_2^2$ since V^{T} is orthogonal, and orthogonal matrices preserve L^2 norm (by definition). Thus defining $\beta^V = V^{\mathsf{T}} \beta$ and $z = U^{\mathsf{T}} y$, we wish to solve

$$\begin{aligned}
&\text{minimize } \left\|\beta^V\right\|_2^2 \\
&\text{subject to } \Lambda \tilde{\beta} = \Sigma^{\mathsf{T}} z.
\end{aligned}$$

Due to the diagonal form of Λ, this turns out to be an easy problem. Observe that $(\Sigma^{\mathsf{T}} z)_i = 0$ for $i > r$, since $\Sigma_{ii} = 0$ for $i > r$. Thus, in block matrix form, we have

$$\left[\begin{array}{ccc|c}
\lambda_1 & & & \\
& \ddots & & \\
& & \lambda_r & \\
\hline
& & & 0_{(d-r)\times(d-r)}
\end{array}\right] \cdot \left[\begin{array}{c}
\beta_1^V \\
\vdots \\
\beta_r^V \\
\hline
\beta^V_{r+1,\dots,d}
\end{array}\right] = \left[\begin{array}{ccc|c}
\sigma_1 & & & \\
& \ddots & & \\
& & \sigma_r & \\
\hline
& & & 0_{(d-r)\times(n-r)}
\end{array}\right] \cdot z$$

where $0_{p\times q}$ are all-zero matrices of size $p \times q$. It is now clear that the minimal β^V must set $\beta^V_{r+1} = \ldots = \beta^V_d = 0$, and that the remaining β^V are given by

$$\left[\begin{array}{c} \beta^V_1 \\ \vdots \\ \beta^V_r \\ \hline \beta^V_{r+1,\ldots,d} \end{array}\right] = \left[\begin{array}{ccc|c} \sigma_1^{-1} & & & \\ & \ddots & & \\ & & \sigma_r^{-1} & \\ \hline & & & 0_{(d-r)\times(n-r)} \end{array}\right] \cdot z$$

Restoring $\beta^V = V^\top \beta$ and $z = U^\top y$, this is the same as

$$\beta = V \cdot \left[\begin{array}{ccc|c} \sigma_1^{-1} & & & \\ & \ddots & & \\ & & \sigma_r^{-1} & \\ \hline & & & 0_{(d-r)\times(n-r)} \end{array}\right] \cdot U^\top y.$$

This matrix we have on the right is exactly the same matrix as Σ^+ defined earlier, and in fact, this minimal L^2 solution is the same as the $\lim_{t\to\infty} \beta_t$ we derived at the beginning of this section! The matrix Σ^+ is called the *Moore-Penrose pseudoinverse* of Σ.

All our computations thus far show that as $t \to \infty$, gradient flow converges not only to a solution of the least squares problem, but also, *implicitly regularizes* the solution to have the least possible L^2 norm (see Figure 1). This is an observation of great importance, since it argues that not only is gradient descent just a method which finds *some* solution, but it also, implicitly, chooses a "low complexity" solution. Of course, this property is only active when the solution is not unique (like when $r < d$ in our example), but this is indeed the case in many modern applications, such as neural networks, which are *extremely overparameterized*.

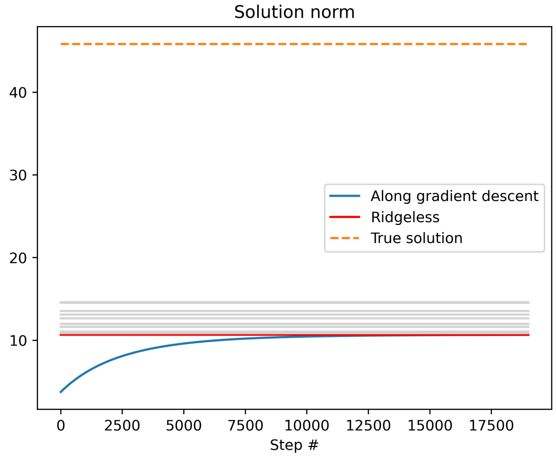

Figure 1: Implicit regularization of gradient flow. The gray lines are other solutions with *zero loss* but higher L^2 norm than the ridgeless solution (shown in red).

4 Are norms always the answer?

The most common form of regularization studied in the literature is *norm-based regularization*. For instance, in LASSO and Ridge, we regularize by the L^1 and L^2 norms of the parameter vector β, explicitly, via a tunable penalty strength. As we saw above, even the *implicit regularization* of gradient flow is norm-based, in particular, L^2-norm based. One may ask if this is a general phenomenon, i.e., is gradient flow always regularizing via *some norm*?

In the paper [5], the authors argue the opposite, exhibiting a class of *matrix factorization* problems for which gradient flow minimizes *no norm* (in the sense of vector norms, with matrices thought of as vectors), but instead minimizes a notion of *rank*. The rest of this document is devoted to describing the results established there.

5 Deep matrix factorization

The problem under consideration is referred to as *deep matrix factorization*. Here, we would like to complete a $d \times d'$ matrix given some of its entries. Say $\Omega \subseteq [d] \times [d']$ (recall that $[n] = \{1, \dots, n\}$) are the indices of the known entries $\{b_{ij}\}_{(i,j)\in\Omega}$. Given a candidate matrix W, the loss ℓ is defined as

$$\ell(W) = \frac{1}{2} \sum_{(i,j)\in\Omega} (W_{ij} - b_{ij})^2 . \tag{7}$$

The strategy employed is the following. We will find L matrices $W_1, \dots, W_L$ of hidden dimensions $d_1, \dots, d_{L-1}$ such that $W_i \in \mathbb{R}^{d_i \times d_{i-1}}$ (defining $d_L \overset{\text{def}}{=} d, d_0 \overset{\text{def}}{=} d'$), and use $W_{L:1}$ as our candidate completion W where

$$W_{L:1} \overset{\text{def}}{=} W_L W_{L-1} \cdots W_1. \tag{8}$$

Our goal now is to understand the kind of matrices W_i found by gradient flow when run on this problem with loss ℓ.

A particularly simple instance of this problem, as described in [5], captures the main thesis quite beautifully, which we present below. The general arguments are similar to this simplified case, but require greater technical effort, so for the purposes of brevity and clarity, we stick to this case.

Suppose $d = d' = 2$, so that we are completing 2×2 matrices, and suppose everything except the $(1, 1)$ entry is known and have the following values:

$$b = \begin{bmatrix} * & 1 \\ 1 & 0 \end{bmatrix} \tag{9}$$

(here $*$ is the unknown entry). Define the set of matrices with *zero loss* as

$$\mathcal{S} \overset{\text{def}}{=} \{W \in \mathbb{R}^{2\times 2} : W_{ij} = b_{ij}, (i,j) \in \Omega\}. \tag{10}$$

It is clear that for any norm $\|\cdot\|$, there is a $K > 0$ such that any minimizer of $\|W\|$ with $W \in \mathcal{S}$ must have $|W_{11}| \le K$. In fact, this can also be upgraded to ϵ-minimizers, by replacing K with an ϵ-dependent K_ϵ. To see why, observe that any $W \in \mathcal{S}$ with $|W_{11}| \ge t$ must have $\|W\| \ge c_1 t - c_2$ for constants c_1, c_2 depending on $\|\cdot\|$ via the triangle inequality and scaling for norms. Since $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \in \mathcal{S}$ and has finite norm, we have our claim. In summary, *optimizing norms requires the method to keep W_{11} bounded.*

However, as we will see now, bounded W_{11} is in *direct contradiction* with *rank minimization*. Each matrix in $\mathcal{S}$ is of rank 2, but, heuristically, as $W_{11} \to \infty$, the matrix is *effectively of rank 1*, since the other entries are much smaller compared to W_{11}. Of course, this is not rigorously true for the usual discrete rank, so, following the authors, we will replace it with two other notions of rank that capture this "*essentially rank 1*" notion:

- **Effective rank** (of a matrix $0 \ne W \in \mathbb{R}^{d\times d'}$): Let $\sigma_1, \dots, \sigma_r$ be the singular values of W where r is the (usual) rank of W. The *effective rank* is

  $$\operatorname{erank}(W) \overset{\text{def}}{=} \exp\left(H(\rho_1, \dots, \rho_r)\right),$$

  where $H(\lambda_1, \dots, \lambda_r) = -\sum_k \lambda_k \ln \lambda_k$ is the Shannon entropy (with $0 \ln 0 = 0$) and ρ are the normalized singular values

  $$\rho_i \overset{\text{def}}{=} \frac{\sigma_i}{\sum_{j=1}^{r} \sigma_j}, \quad i = 1, \dots, r.$$

For instance, for the identity matrix I_d, $\sigma_i = 1$ for all i, so that $\rho_i = 1/d$. Thus, the Shannon entropy (of this uniform distribution) is $\ln d$, and the effective rank $= \exp(\ln d) = d$, which makes sense.

Further, if one entry W_{11} is too high, for instance if $W = 10^3 E_{11} + I_d$ (where E_{11} is the all zeros matrix except the $(1,1)$ entry which is 1), then the largest singular value is of the order of 10^3, and everything else is much smaller. Hence the distribution ρ_i is very singular, and the Shannon entropy is close to 0, making the effective rank 1, which justifies why we would call this the "effective rank".

- **Infimal rank** (of a set of matrices $\mathcal{S} \subseteq \mathbb{R}^{d\times d'}$): For any two sets $\mathcal{S}, \mathcal{S}' \subseteq \mathbb{R}^{d\times d'}$, the *Frobenius distance* between these two sets is

$$D(\mathcal{S}, \mathcal{S}') \overset{\text{def}}{=} \inf\{\|W - W'\|_F : W \in \mathcal{S}, W' \in \mathcal{S}'\},$$

where $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm. The *infimal rank* of the set $\mathcal{S}$, $\operatorname{irank}(\mathcal{S})$, is then the smallest r such that $D(\mathcal{S}, \mathcal{M}_r) = 0$ where $\mathcal{M}_r \subseteq \mathbb{R}^{d\times d'}$ are all matrices of rank at most r. Further, the *distance of a matrix W from the infimal rank of $\mathcal{S}$* is defined to be $D(W, \mathcal{M}_{\operatorname{irank}(\mathcal{S})}) \overset{\text{def}}{=} D(\{W\}, \mathcal{M}_{\operatorname{irank}(\mathcal{S})})$.

For the set $\mathcal{S}$ defined in (10), we have $\operatorname{irank}(\mathcal{S}) = 1$ (and not 2). To see why, observe that for any $a \geq 0$,

$$W \overset{\text{def}}{=} \begin{bmatrix} a & 1 \\ 1 & 0 \end{bmatrix} = [a^{1/2}, a^{-1/2}]^\top [a^{1/2}, a^{-1/2}] - a^{-1} E_{22},$$

so the Frobenius distance between the matrix on the left and the first term on the right is a^{-1} (which is rank 1), goes to 0 as $a \to \infty$. This also shows that $D(W, \mathcal{M}_1) \leq a^{-1}$.

The main theorem is the following (Theorem 1 in [5], simplified for presentation)

Theorem 1. *Suppose we minimize the loss in (7) over an L-depth matrix factorization as in (8) using gradient flow. Denote by $W_{1:L}(t)$ the product in (8) at time t along the flow, and by $\ell(t)$ the corresponding loss. Further suppose at that initialization, $\det(W_{1:L}(0)) > 0$. Then for any norm $\|\cdot\|$ (see Figure 2)*

$$\|W_{1:L}(t)\| \geq \frac{a}{\sqrt{\ell(t)}} - b, \quad \forall t \geq 0, \tag{11}$$

where a, b are constants depending only on the norm $\|\cdot\|$. Further, for the two notions of rank defined above, we have (see Figure 3)

$$\operatorname{erank}(W_{1:L}(t)) \leq \inf_{W' \in \mathcal{S}} \operatorname{erank}(W') + c_1\sqrt{\ell(t)}, \tag{12}$$

$$D(W_{1:L}(t), \mathcal{M}_{\operatorname{irank}(\mathcal{S})}) \leq c_2\sqrt{\ell(t)}, \quad \forall t \geq 0 \tag{13}$$

where c_1, c_2 are two universal constants.

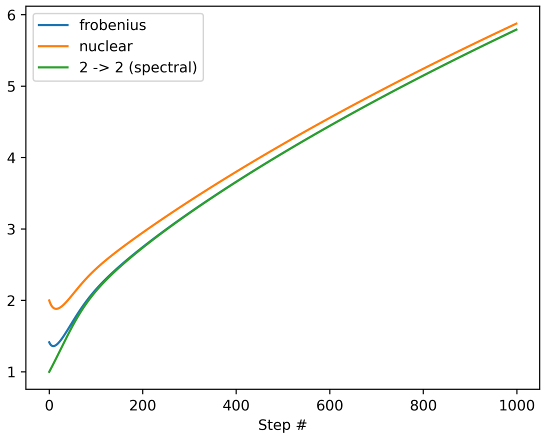

Figure 2: Frobenius, nuclear and spectral norms are all sent to infinity along gradient flow.

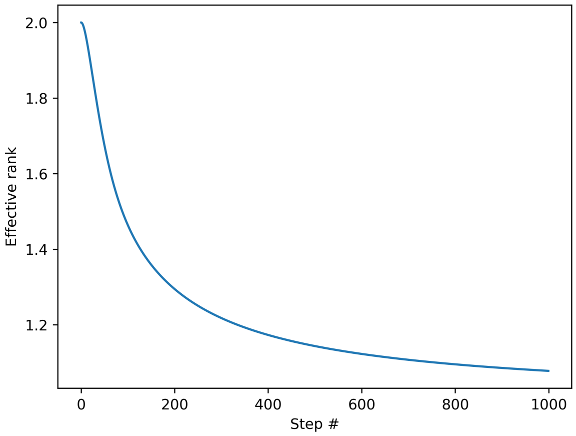

Figure 3: Ranks are minimized along gradient flow.

Although we will not prove it in this report, under certain reasonable initializations, gradient flow indeed sends losses to zero (observe that zero loss is always achievable), see for instance Proposition 4 in [5]. In that case, (11) shows that the norms are sent to infinity. On the contrary, (12) shows that effective rank gets close to the best it can be and that the flow selects the one with the smallest erank, approximately (and exactly as $t \to \infty$). A similar conclusion is implied for irank by (13).

6 Gradient flow for deep matrix factorization

Before we start with the analysis required for Theorem 1, we will first need to understand gradient flow in this case. The results here are from [1], which is also cited by our primary source [5], and forms a precursor to the results of this paper.

In particular, the result we need is Theorem 1 from [1],

Theorem 2 (Gradient flow for matrix factorization)**.** *Suppose we run gradient flow over the factorization in* (8)*, i.e., the evolution of the matrices $W_j, j = 1, \dots, L$ are given by*

$$\dot{W_j}(t) \overset{\mathrm{def}}{=} \frac{d}{dt} W_j(t) = -\frac{\partial}{\partial W_j} \ell(W_L W_{L-1} \cdots W_1),$$

(here $\frac{\partial}{\partial W_j}$ means the partial derivative w.r.t. each entry of W_j, arranged as a matrix), with **balanced initialization***,*

$$W_{j+1}^{\mathsf{T}}(0) W_{j+1}(0) = W_j(0) W_j^{\mathsf{T}}(0), \quad \forall j = 1, \dots, L-1.$$

(see discussion below). Then the product matrix $W_{1:L}(t)$ follows the dynamics

$$(\dot{W_{1:L}})(t) = -\sum_{j=1}^{L} \left[W_{1:L}(t) W_{1:L}(t)^{\mathsf{T}} \right]^{\frac{j-1}{L}} \cdot \nabla \ell(W_{1:L}(t)) \cdot \left[W_{1:L}(t)^{\mathsf{T}} W_{1:L}(t) \right]^{\frac{L-j}{L}} \tag{14}$$

where $[A]^{\alpha}$ is the α-th power of a PSD matrix A, defined naturally via its spectral decomposition, for $\alpha \geq 0$.

The condition of *balancedness* simplifies the results. It is also, approximately, satisfied for random near-zero initializations. We do not discuss this point further, but refer to Section 3 in [5] for a deeper discussion. However, a basic point we wish to mention is that the *all-zero* initialization is not valid, since all the partial derivatives are zero in that case, preventing gradient flow from any movement. Near-zero initializations are used to circumvent this issue.

The dynamics described in Theorem 2 is cleaner and more tractable in terms of the SVD of $W_{1:L}(t)$. However, producing singular value dynamics from matrix dynamics requires some smoothness along time t of the SVD. Fortunately, as invoked in [2], there exist *analytic singular value decompositions* via classical results of [3], which in turn follows from an *analytic spectral decomposition* (see for instance, the book by Kato [4], Theorem 6.1). *However, for such a result, one must give up the usual nonnegativeness of the singular values and the assumption that the singular values are nonincreasing. See [3], Section 2.2 for an example illustrating this point.*

For our purposes, the following result from [2] is sufficient:

Lemma 3. *In the notation above, let $m = \min(d, d')$. Then, there are matrices $U \in \mathbb{R}^{d\times m}, V \in \mathbb{R}^{d'\times m}$ with orthonormal columns, and a matrix of singular values $S(t) \in \mathbb{R}^{m\times m}$ which is diagonal (but not necessarily nonnegative and nonincreasing along the diagonal) such that*

$$W_{1:L}(t) = U(t)S(t)V(t)^{\mathsf{T}} \tag{15}$$

for all $t \geq 0$.

The proof of this lemma, in view of [3] Theorem 1, only requires showing that $W_{1:L}(t)$ is analytic, which in turn, follows directly from classical facts about ODEs and the analyticity of the loss $\ell(W_L W_{L-1}\cdots W_1)$ as functions of the matrices W_i. Crucially, the analytic form of the loss is important, and won't hold for more singular loss functions.

Armed with Lemma 3, one may substitute the decomposition (15) into Theorem 2 to obtain

Lemma 4. *The dynamics of the analytic singular values in (15) are given by*

$$\dot{\sigma}_j(t) = -L\cdot(\sigma_j^2(t))^{1-1/L}\left\langle \nabla\ell(W(t)), u_j(t)v_j(t)^{\mathsf{T}}\right\rangle, \quad j = 1,\dots,m. \tag{16}$$

Proof. Differentiating the SVD (15) we get (shortening $W_{1:L}$ to W for the moment)

$$\dot{W}(t) = \dot{U}(t)S(t)V(t)^{\mathsf{T}} + U\dot{S}(t)V(t)^{\mathsf{T}} + US(t)\dot{V}(t)^{\mathsf{T}}.$$

Premultiplying by $U(t)^{\mathsf{T}}$ and postmultiplying by $V(t)$ produces

$$U(t)^{\mathsf{T}}\dot{W}(t)V(t) = U(t)^{\mathsf{T}}\dot{U}(t)S(t) + \dot{S}(t) + S(t)\dot{V}(t)^{\mathsf{T}}V(t)$$

using the orthonormality of the columns (and $m \leq d$). Looking at the diagonal elements of this equation, we get

$$u_j(t)^{\mathsf{T}}\dot{W}(t)v_j(t) = \langle u_j(t), \dot{u}_j(t)\rangle\,\sigma_j(t) + \dot{\sigma}_j(t) + \sigma_j(t)\,\langle \dot{v}_j(t), v_j(t)\rangle\,, \quad j = 1,\dots,m.$$

where $\sigma_j(t)$ are the diagonal elements of $S(t)$, and $u_j(t), v_j(t)$ are the columns of $U(t), V(t)$ respectively. Importantly, observe that since $u_j(t)$ always has length 1,

$$0 = \frac{d}{dt}\,\|u(t)\|_2^2 = 2\,\langle u_j(t), \dot{u}_j(t)\rangle\,,$$

and so the equation above reduces to

$$\dot{\sigma}_j(t) = u_j(t)^{\mathsf{T}}\dot{W}(t)v_j(t). \tag{17}$$

Now we plug in the analytic SVD (15) into the dynamics (14) for $W(t)$ to obtain

$$\begin{aligned}\dot{W}(t) &= -\sum_{j=1}^{L}\left[W(t)W(t)^{\mathsf{T}}\right]^{\frac{j-1}{L}}\cdot\nabla\ell(W(t))\cdot\left[W(t)^{\mathsf{T}}W(t)\right]^{\frac{L-j}{L}}\\ &= -\sum_{j=1}^{L}U(t)(S^2(t))^{\frac{j-1}{L}}U(t)^{\mathsf{T}}\cdot\nabla\ell(W(t))\cdot V(t)(S^2(t))^{\frac{L-j}{L}}V(t)^{\mathsf{T}}\end{aligned}$$

In view of (17), we premultiply the last equation with $u_j(t)^{\mathsf{T}}$ and postmultiply by $v_j(t)$ and once again use the orthonormality of the columns of U, V to get

$$\begin{aligned}\dot{\sigma}_j(t) &= -u_j(t)^{\mathsf{T}}\left(\sum_{j=1}^{L}U(t)(S^2(t))^{\frac{j-1}{L}}U(t)^{\mathsf{T}}\cdot\nabla\ell(W(t))\cdot V(t)(S^2(t))^{\frac{L-j}{L}}V(t)^{\mathsf{T}}\right)v_j(t)\\ &= \sum_{j=1}^{L}(\sigma_j^2(t))^{\frac{j-1}{L}}u_j(t)^{\mathsf{T}}\cdot\nabla\ell(W(t))\cdot(\sigma_j^2(t))^{\frac{L-j}{L}}v_j(t)\\ &= -L\cdot(\sigma_j^2(t))^{1-1/L}\left\langle\nabla\ell(W(t)), u_j(t)v_j(t)^{\mathsf{T}}\right\rangle\end{aligned}$$

as required. □

7 Blowup of norms

A key ingredient of our proof of the blowup of norms (11) is the following observation.

Lemma 5. *We continue with the notation $W(t) = W_{1:L}(t)$. Assume that the depth $L \geq 2$, and $d = d'$ (so that we include our simple example (9)). Then if $\det(W(0)) > 0$, we have $\det(W(t)) > 0$ throughout as well.*

Proof. Observe that if $\sigma_j(0) = 0$, then $\sigma_j(t) = 0$ for all t as well, as is evident from the dynamics (16) when $L \geq 2$. Similarly, we will establish a *sign preservation property*, i.e., if $\sigma_j(0) < 0$ then $\sigma_j(t) < 0$ throughout, and analogously, for $\sigma_j(0) > 0$. This will finish the proof since the determinant is the product of the singular values.

Fix an index j. Let $g(t) \overset{\text{def}}{=} -L \left\langle \nabla \ell(W(t), u_j(t) v_j(t)^\mathsf{T} \right\rangle$, which is some analytic function since W, U, V, ℓ are all analytic functions. Then the differential equation in (16) is just

$$\dot{\sigma}_j(t) = g(t)(\sigma_j(t))^{2\alpha}, \quad \alpha = 2(1 - 1/L) \geq 1.$$

If $\alpha = 1$, i.e., $\dot{\sigma}_j(t) = g(t)\sigma_j(t)$, the solution is given by $\sigma_j(t) = \sigma_j(0) \exp\left(\int_0^t g(s) ds\right)$ (as can be seen by differentiation), and hence $\sigma_j(0)$ has the same sign as $\sigma_j(t)$. If, on the other hand, $\alpha > 1$, first assume $\sigma_j(0) > 0$. Observe that $\frac{d}{dt}\left(\sigma_j(t)^{1-2\alpha}\right) = (1-2\alpha)\dot{\sigma}_j(t)\sigma_j(t)^{-2\alpha}$, and since the above equation is the same as $\dot{\sigma}_j(t)\sigma_j(t)^{-2\alpha} = g(t)$, we have

$$\frac{d}{dt}\left(\frac{1}{1-2\alpha}\sigma_j(t)^{1-2\alpha}\right) = g(t).$$

The solution to this equation is then

$$\sigma_j(t) = \left(\sigma_j(0)^{1-2\alpha} + \int_0^t (1-2\alpha) g(s) ds\right)^{\frac{1}{1-2\alpha}}$$

Observe that since $\sigma_j(0) > 0$, $\sigma_j(t) > 0$ as long as the quantity inside the brackets above is > 0, after which this blows up. Of course, this is for general $g(t)$, but our dynamics guarantees the analyticity of the solution $\sigma_j(t)$, so the latter does not happen.

For the case when $\sigma_j(0) < 0$, simply consider the analogous equation for $-\sigma_j(t)$, establishing the claim. □

Armed with this result, we will prove (11) in this section. First observe that for our problem,

$$\ell(W) = \frac{1}{2}\left[(W_{12} - 1)^2 + (W_{21} - 1)^2 + W_{22}^2\right],$$

so that

$$|W_{12} - 1|, |W_{21} - 1| \leq \sqrt{2\ell(W)}, \tag{18}$$
$$|W_{22}| \leq \sqrt{2\ell(W)}. \tag{19}$$

Denoting $W_{1:L}$ by W, we also know that $\det W(t) > 0$ throughout since $\det W(0) > 0$ by assumption, which implies

$$0 < \det W(t) = W_{11}(t)W_{22}(t) - W_{12}(t)W_{21}(t).$$

Now suppose $\ell(W) < 1/2$. Then from (18) we see that $W_{12}, W_{21} \in (0, 2)$, and thus

$$|W_{11}W_{22}| > |W_{12}W_{21}| = |W_{12}|\,|W_{21}| \overset{(18)}{\geq} \left(1 - \sqrt{2\ell(W)}\right)^2.$$

and since $W_{22} \neq 0$, we have

$$|W_{11}| \geq \frac{\left(1 - \sqrt{2\ell(W)}\right)^2}{\sqrt{2\ell(W)}} = \frac{1}{\sqrt{2\ell(W)}} - 2 + \sqrt{2\ell(W)}. \tag{20}$$

Armed with this observation, we first show the result of equation (11). In fact, the only fact we need about our norm $\|\cdot\|$ on matrices, other than homogeneity, is that they satisfy a *quasi*-triangle inequality, i.e.,

$$\|A + B\| \leq c \cdot (\|A\| + \|B\|), \quad \forall A, B \tag{21}$$

for some fixed constant $c \geq 1$ (in particular, the results also hold for *quasi*-norms). Note that this is just the usual triangle inequality when $c = 1$. The key idea is that since $W_{1,1} \to \infty$ while $W_{1,2}, W_{2,1}, W_{2,2}$ are all bounded due to loss going to zero, the overall norm must go to infinity just via triangle-inequality-like bounds.

Formally, in order to bound $\|W_{1:L}\|$, we apply equation (21) to $\left\|W_{1,1}e_1e_1^\top\right\|$, which produces

$$\left\|W_{1,1}e_1e_1^\top\right\| = \left\|W_{1,1}e_1e_1^\top - W_{1:L} + W_{1:L}\right\| \leq c \cdot \left(\left\|W_{1,1}e_1e_1^\top - W_{1:L}\right\| + \|W_{1:L}\|\right).$$

Here, e_1 and e_2 refer to the standard basis of 2×2 matrices. Rearranging the above yields

$$\|W_{1:L}\| \geq \frac{1}{c}\left\|W_{1,1}e_1e_1^\top\right\| - \left\|W_{1,1}e_1e_1^\top - W_{1:L}\right\|.$$

Now we bound both terms on the right side. Homogeneity for norms states that $\|\alpha A\| = |\alpha| \cdot \|A\|$ holds for all scalars $\alpha \in \mathbb{R}$ and matrices A. Using this fact we may bound:

$$\begin{aligned}\left\|W_{1,1}e_1e_1^\top - W_{1:L}\right\| &= \left\|\cancel{W_{1,1}e_1e_1^\top} - \left(W_{1,2}e_1e_2^\top + W_{2,1}e_2e_1^\top + W_{2,2}e_2e_2^\top + \cancel{W_{1,1}e_1e_1^\top}\right)\right\| \\ &= \left\|W_{1,2}e_1e_2^\top + W_{2,1}e_2e_1^\top + W_{2,2}e_2e_2^\top\right\|.\end{aligned}$$

Applying the quasi-triangle inequality (21) twice we continue with

$$\begin{aligned}&\leq c \cdot \left(\left\|W_{1,2}e_1e_2^\top\right\| + c \cdot \left(\left\|W_{2,1}e_2e_1^\top\right\| + \left\|W_{2,2}e_2e_2^\top\right\|\right)\right) \\ &= c\,|W_{1,2}| \cdot \left\|e_1e_2^\top\right\| + c^2\,|W_{21}| \cdot \left\|e_2e_1^\top\right\| + c^2\,|W_{22}| \cdot \left\|e_2e_2^\top\right\|.\end{aligned}$$

Now invoking (18) and (19) we get

$$\begin{aligned}&\leq \max\left\{c(1 + \sqrt{2l(W)}), c^2(1 + \sqrt{2l(W)}), c^2\sqrt{2l(W)}\right\} \cdot \left(\left\|e_1e_2^\top\right\| + \left\|e_2e_1^\top\right\| + \left\|e_2e_2^\top\right\|\right) \\ &\leq K(1 + \sqrt{\ell(W)}),\end{aligned}$$

where K is a constant depending only the norm $\|\cdot\|$. This quantity is bounded since gradient flow only decreases the loss. On the contrary, the remaining term is

$$\frac{1}{c}\left\|W_{1,1}e_1e_1^\top\right\| = \frac{1}{c}|W_{1,1}| \cdot \left\|e_1e_1^\top\right\| \overset{(20)}{\geq} \frac{1}{c} \cdot \left(\frac{1}{\sqrt{2l(W)}} - (2 - \sqrt{2l(W)})\right)\left\|e_1e_1^\top\right\| \geq \frac{K'}{\sqrt{\ell(W)}} - K''$$

where K', K'' are again two constants depending only on the norm.

Combining the two bounds above, we then have

$$\|W_{1:L}\| \geq \frac{K'}{\sqrt{\ell(W)}} - K'' - K\left(1 + \sqrt{\ell(W)}\right) \geq \frac{a}{\sqrt{\ell(W)}} - b,$$

for constants a, b depending only on the norm (and a bound on the initial loss), finishing the proof of (11). This shows that all norms blow up as the loss is sent to zero.

8 Rank minimization

The key observation here is that since one entry is sent to infinity, one singular value of W becomes much larger than the rest, making the matrix effectively rank 1. To prove this rigorously, we will first prove some simple singular value bounds for general 2×2 matrices. In these bounds, we always think of A_{11} as being large compared to the remaining entries. *Note that our proofs here are softer variants of the arguments in the paper, which we choose for simplicity and clarity purposes.*

Lemma 6. *Let $A \in \mathbb{R}^{2\times 2}$ be the following matrix*

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}.$$

Suppose $\sigma_1 \geq \sigma_2$ are the two singular values of A. Then, $\sigma_1^2 \geq a^2 + c^2$ and $\sigma_1^2 + \sigma_2^2 = a^2 + b^2 + c^2 + d^2$. Thus, $\sigma_2^2 \leq b^2 + d^2$.

Proof. Since

$$A^\mathsf{T} A = \begin{bmatrix} a^2 + c^2 & ab + cd \\ ab + cd & b^2 + d^2 \end{bmatrix},$$

we have

$$\sigma_1^2 + \sigma_2^2 = \operatorname{tr}(A^\mathsf{T} A) = a^2 + b^2 + c^2 + d^2.$$

Also by the variational problem for eigenvalues,

$$\sigma_1^2 = \sup_{\|u\|_2 = 1} u^\mathsf{T} A^\mathsf{T} A u.$$

Choosing $u = [1, 0]^\mathsf{T}$, we get $\sigma_1^2 \geq a^2 + c^2$. □

Applying Lemma 6 to W we have

$$\begin{aligned} \sigma_1(W) &\geq \sqrt{W_{1,1}^2 + W_{2,1}^2} \\ &\geq |W_{1,1}| \\ &\overset{(20)}{\geq} \frac{1}{\sqrt{2\ell(W)}} - 2, \end{aligned}$$

meanwhile

$$\begin{aligned} \sigma_2(W) &\leq \sqrt{W_{1,2}^2 + W_{2,2}^2} \\ &\overset{(18)}{\leq} \sqrt{2\left(1 + 2\sqrt{\ell(W)}\right)^2} \\ &\leq c_1 + c_2\sqrt{\ell(W)} \end{aligned}$$

for absolute constants c_1, c_2. Thus the ρ_2 in the definition of erank satisfies

$$\rho_2 = \frac{\sigma_2}{\sigma_1 + \sigma_2} \leq \frac{c_1 + c_2\sqrt{\ell(W)}}{c_1 + c_2\sqrt{\ell(W)} + \frac{1}{\sqrt{2\ell(W)}} - 2} \lesssim \sqrt{\ell(W)}$$

as long as $\ell(W) \to 0$ (where c_1', c_2' are other constants). Observing that

$$H(1-\epsilon, \epsilon) = -(1-\epsilon)\log(1-\epsilon) - \epsilon \log \epsilon \lesssim (1-\epsilon)\epsilon \lesssim \epsilon$$

as $\epsilon \to 0$, we then have

$$\operatorname{erank}(W) = \exp(H(\rho_1, \rho_2)) = \exp(O(\rho_2)) \leq 1 + O(\sqrt{\ell(W)})$$

where the $O(\cdot)$ hides absolute constants as $\ell(W) \to 0$. Compare this to (12). Clearly, the smallest rank a zero-loss solution in $\mathcal{S}$ could have is 1, so this is the required inequality.

Turning to infimal rank, recall that we described earlier why $\operatorname{irank}(\mathcal{S}) = 1$. Note that if

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \quad a > 0,$$

is such that $a \gg b, c, d$, the rank 1 matrix $A' = uu^\top$ with $u = [a^{1/2}, a^{-1/2}]^\top$ is close to A in Frobenius norm, i.e.,

$$\|A - A'\|_F^2 = \left\| \begin{bmatrix} 0 & b-1 \\ c-1 & d-a^{-1} \end{bmatrix} \right\|_F^2 = (b-1)^2 + (c-1)^2 + (d-a^{-1})^2$$

If $A = W$, the latter expression is very similar to $\ell(W)$ except for the term $(d - a^{-1})^2$ instead of $(d-0)^2$. But our prior arguments show that $a = W_{1,1}$ is huge, so this term is also similar. However as loss $\ell(W)$ is sent to zero, Frobenius norm is also sent to 0, and thus the distance to infimal rank for W goes to 0 as claimed in (13). Of course, these are heuristic arguments without precise estimates as in (13), but the formal arguments are similar to the erank case and are left out for brevity.

9 Conclusions

Gradient descent, despite its virtues, remains an elusive source of implicit regularization, sometimes choosing rank over norm optimization for overparameterized problems. The core goal of this report was to illustrate the case of deep matrix factorization, wherein smoothed variants of rank are preferred by gradient flow, sending entries of the matrix to infinity, in complete defiance of norm-regularization. We present many proofs from [5] supporting the claims (some of which are technical in nature), but we hope that the message is clear: *rank should be considered as a potential candidate for implicit regularization by gradient flow in real-world problems.* Simulations presented in our source [5] also indicate similar aspects in tensor factorization, further emphasizing the ubiquity of this seemingly pathological behavior.

References

[1] Sanjeev Arora, Nadav Cohen, and Elad Hazan. “On the optimization of deep networks: Implicit acceleration by overparameterization”. In: *International conference on machine learning*. PMLR. 2018, pp. 244–253.

[2] Sanjeev Arora et al. “Implicit regularization in deep matrix factorization”. In: *Advances in Neural Information Processing Systems* 32 (2019).

[3] Angelika Bunse-Gerstner et al. “Numerical computation of an analytic singular value decomposition of a matrix valued function”. In: *Numerische Mathematik* 60 (1991), pp. 1–39.

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

[5] Noam Razin and Nadav Cohen. “Implicit regularization in deep learning may not be explainable by norms”. In: *Advances in neural information processing systems* 33 (2020), pp. 21174–21187.