Gradient Flow Implicitly Regularizes to the Minimum-Norm Solution in Overparameterized Linear Regression

Jatin Batra

August 28, 2025

Abstract

In modern machine learning, it has been observed that gradient-based optimization algorithms tend to find *implicit* forms of regularization, favoring solutions with certain desirable properties even in the absence of explicit regularizers. In the context of linear regression with more parameters than data points (overparameterization), gradient descent is known to converge to the solution with minimum Euclidean norm among all interpolating solutions[1]. This document provides a rigorous, pedagogical proof of this phenomenon using continuous-time gradient flow. We begin by reviewing the necessary linear algebra background and the definition of the Moore-Penrose pseudoinverse, which gives the minimum-norm solution to linear systems[2]. We then set up the gradient flow for overparameterized linear regression and state the main theorem: starting from zero initial conditions, the gradient flow converges to the unique minimum-norm solution that perfectly fits the data. A detailed proof is presented, accompanied by intuitive commentary. Finally, we discuss the interpretation of this result and provide a few exercises to solidify understanding.

Contents

[1] https://math.stackexchange.com

[2] https://en.wikipedia.org

1 Introduction

In linear regression, we seek to fit a linear model to a given set of data. Consider a dataset of m examples $(x_i, y_i)_{i=1}^m$, where each $x_i \in \mathbb{R}^n$ is a feature vector and $y_i \in \mathbb{R}$ is a target value. The linear model assumes a relationship $y_i \approx w^\top x_i$, where $w \in \mathbb{R}^n$ is a vector of parameters or weights. The goal is to find a w that produces predictions $w^\top x_i$ as close as possible to y_i for all i. Equivalently, in matrix-vector form, we have:

Xw ≈ y,

where X is the $m \times n$ design matrix whose i-th row is $x_i^\top$, and $y \in \mathbb{R}^m$ is the vector of targets.

A common approach is to choose w by minimizing the *least-squares* cost:

$$L(w) = \frac{1}{2}\|Xw - y\|_2^2 = \frac{1}{2}\sum_{i=1}^{m}(w^\top x_i - y_i)^2.$$

This is a convex quadratic optimization problem. When X has full column rank (i.e. $n \leq m$ and the columns of X are independent), the solution is unique and given by the normal equations $X^\top X w = X^\top y$. However, in modern practice it is common to encounter *overparameterized* settings where $n > m$ (more parameters than data points). In this case, X does not have full column rank; instead, it typically has *full row rank* (rank m), assuming the m data points are independent. When $\operatorname{rank}(X) = m < n$, the linear system $Xw = y$ is underdetermined: there are infinitely many solutions (assuming the system is consistent, i.e. y lies in the column space of X).

A natural question arises: **which** of the infinitely many minimizers of $L(w)$ (all of which satisfy $Xw = y$) is found by a given optimization procedure? In particular, it has been empirically observed and theoretically proven that plain gradient descent (with suitable initialization) on the unregularized least-squares cost converges to the *minimum Euclidean norm* solution w that interpolates the data[3]. This is often described by saying that gradient descent *implicitly regularizes* towards the minimum-norm solution, even though no explicit norm penalty is present in $L(w)$.

In these notes, we will prove this fact rigorously in the simplified setting of *continuous-time gradient flow*. Gradient flow can be thought of as the limit of gradient descent as the step size tends to zero, yielding a differential equation that the parameters follow. By analyzing this differential equation, we can precisely characterize the limiting behavior of the parameters $w(t)$ as $t \to \infty$ and show that $w(t)$ converges to the minimum-norm interpolating solution.

We proceed as follows. In Section 2, we review some mathematical preliminaries in linear algebra, including the fundamental subspaces associated with a matrix (column space, nullspace, row space) and orthogonal projections. In Section 3, we introduce the Moore-Penrose pseudoinverse, which provides a closed-form expression for the minimum-norm solution of $Xw = y$, and prove its key properties. In Section 4, we set up the gradient flow dynamic for linear regression and precisely define what we mean by an "overparameterized" regime. Section 5 contains the statement of the main theorem on the convergence of gradient flow to the minimum-norm solution. Section 6 provides a full detailed proof of the theorem

[3] `https://math.stackexchange.com`; `https://www.cs.ubc.ca`

with commentary at each step. We then offer some intuitive interpretation of the result in Section 7, explaining why gradient flow gravitates to this special solution. Finally, Section 8 concludes with a few exercises to test understanding.

2 Mathematical Preliminaries

In this section, we establish some basic linear algebra concepts and results that will be used later. The reader is assumed to have some familiarity with vectors, matrices, and operations like transposition and the dot product, but we will define concepts such as subspaces and projections for completeness.

2.1 Vectors, Norms, and Dot Products

We work in the Euclidean space $\mathbb{R}^n$. We write vectors as bold lower-case letters (e.g. $\mathbf{v} \in \mathbb{R}^n$), and matrices as upper-case letters (e.g. $A \in \mathbb{R}^{m \times n}$). The *dot product* or *inner product* of two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ is

$$u \cdot v = \sum_{i=1}^{n} u_i v_i.$$

The dot product induces the Euclidean *norm* (length) of a vector:

$$\|v\|_2 = v \cdot v = \sum_{i=1}^{n} v_i^2.$$

We will often drop the subscript and write $|\mathbf{v}|$ for $|\mathbf{v}|_2$ when the context is clear.

2.2 Subspaces: Column Space, Nullspace, and Row Space

Given a matrix $A \in \mathbb{R}^{m \times n}$:

- The **column space** (or **range**) of A, denoted $\mathrm{Col}(A)$ or $\mathrm{Range}(A)$, is the set of all vectors in $\mathbb{R}^m$ that can be expressed as $A\mathbf{w}$ for some $\mathbf{w} \in \mathbb{R}^n$. Equivalently, it is the subspace of $\mathbb{R}^m$ spanned by the columns of A.
- The **nullspace** (or **kernel**) of A, denoted $\mathrm{Null}(A)$ or $\ker(A)$, is the set of vectors $\mathbf{v} \in \mathbb{R}^n$ such that $A\mathbf{v} = \mathbf{0}$. It is a subspace of $\mathbb{R}^n$ consisting of all solutions to the homogeneous equation $A\mathbf{v} = 0$.
- The **row space** of A is the column space of $A^\top$ (the transpose of A). It is a subspace of $\mathbb{R}^n$ spanned by the row vectors of A. We denote it by $\mathrm{Row}(A)$ or $\mathrm{Col}(A^\top)$.

These subspaces are related by the dimensions: $\dim(\mathrm{Col}(A)) = \dim(\mathrm{Row}(A)) = \mathrm{rank}(A)$, and the fundamental *rank-nullity* theorem: $\dim(\mathrm{Null}(A)) + \dim(\mathrm{Row}(A)) = n$. Intuitively, $\mathrm{Row}(A)$ and $\mathrm{Null}(A)$ together partition $\mathbb{R}^n$ into two complementary subspaces.

Theorem 1 (Fundamental Theorem of Linear Algebra)**.** *For any matrix $A \in \mathbb{R}^{m\times n}$, the row space of A is the orthogonal complement of the nullspace of A. In symbols:*

$$\mathrm{Row}(A) = (\mathrm{Null}(A))^{\perp},$$

where $(\mathrm{Null}(A))^{\perp} = \{\mathbf{u} \in \mathbb{R}^n : \mathbf{u} \cdot \mathbf{v} = 0$ *for all* $\mathbf{v} \in \mathrm{Null}(A)\}$.

Proof. We give a brief proof. Let $\mathbf{u} \in \mathrm{Row}(A)$. Then $\mathbf{u} = A^{\top}\mathbf{p}$ for some $\mathbf{p} \in \mathbb{R}^m$. Take any $\mathbf{v} \in \mathrm{Null}(A)$, so $A\mathbf{v} = 0$. Then

$u \cdot v = (A^{\top}p) \cdot v = p \cdot (Av) = p \cdot 0 = 0.$

Thus $\mathbf{u}$ is orthogonal to every vector in $\mathrm{Null}(A)$, which shows $\mathrm{Row}(A) \subseteq (\mathrm{Null}(A))^{\perp}$.

Conversely, suppose $\mathbf{u} \in (\mathrm{Null}(A))^{\perp}$. We need to show $\mathbf{u} \in \mathrm{Row}(A)$, i.e. $\mathbf{u}$ can be written as $A^{\top}\mathbf{p}$ for some $\mathbf{p}$. Consider the vector $A\mathbf{u} \in \mathbb{R}^m$. We claim that $\mathbf{p} := A\mathbf{u}$ satisfies $A^{\top}\mathbf{p} = \mathbf{u}$. Indeed, for any $\mathbf{v} \in \mathrm{Null}(A)$, we have $\mathbf{v} \cdot (\mathbf{u} - A^{\top}(A\mathbf{u})) = \mathbf{v} \cdot \mathbf{u} - \mathbf{v} \cdot A^{\top}(A\mathbf{u}) = \mathbf{v} \cdot \mathbf{u} - (A\mathbf{v}) \cdot (A\mathbf{u}) = \mathbf{v} \cdot \mathbf{u} - 0 = 0$. Thus $\mathbf{u} - A^{\top}(A\mathbf{u})$ is orthogonal to the nullspace of A, which by the first part of the proof means $\mathbf{u} - A^{\top}(A\mathbf{u}) \in \mathrm{Row}(A)$. But $A^{\top}(A\mathbf{u})$ is clearly in $\mathrm{Row}(A)$, so their difference $\mathbf{u}$ is also in $\mathrm{Row}(A)$. This completes the proof. □

This fundamental result tells us that any vector in $\mathbb{R}^n$ can be uniquely decomposed into a sum of two components, one lying in the row space of A and the other lying in the nullspace of A. In particular, for any $\mathbf{w} \in \mathbb{R}^n$, there exist unique vectors $\mathbf{w}_{\mathrm{Row}} \in \mathrm{Row}(A)$ and $\mathbf{w}_{\mathrm{Null}} \in \mathrm{Null}(A)$ such that

$w = w_{\mathrm{Row}} + w_{\mathrm{Null}},$

with $\mathbf{w}_{\mathrm{Row}} \cdot \mathbf{w}_{\mathrm{Null}} = 0$.

2.3 Orthogonal Projections onto Subspaces

Given any subspace $S \subseteq \mathbb{R}^n$, for any vector $\mathbf{z} \in \mathbb{R}^n$ there is a distinguished vector in S that is "closest" to $\mathbf{z}$ in Euclidean norm. This is the **orthogonal projection** of $\mathbf{z}$ onto S, defined as the unique vector $\Pi_S(\mathbf{z}) \in S$ such that $\mathbf{z} - \Pi_S(\mathbf{z})$ is orthogonal to every vector in S. Geometrically, $\Pi_S(\mathbf{z})$ is the foot of the perpendicular dropped from $\mathbf{z}$ onto the subspace S. The difference $\mathbf{z} - \Pi_S(\mathbf{z})$ lies in $S^{\perp}$ (the orthogonal complement of S).

One way to characterize the projection is by a minimization property:

$$\Pi_S(z) = \arg\min_{s\in S} \|z - s\|_2.$$

This says $\Pi_S(\mathbf{z})$ is the point in S closest to $\mathbf{z}$. Existence and uniqueness of $\Pi_S(\mathbf{z})$ are standard results in linear algebra (which we will not prove here), relying on the fact that the minimization of a strictly convex quadratic function has a unique solution.

The orthogonal projection operator $\Pi_S : \mathbb{R}^n \to S$ is linear, and satisfies $\Pi_S^2 = \Pi_S$ (projecting twice is the same as projecting once) and $\Pi_S^{\top} = \Pi_S$ (it is self-adjoint with respect to the dot product). A matrix P that satisfies $P^2 = P$ and $P^{\top} = P$ is called an **orthogonal projection matrix**. If P is an orthogonal projection matrix onto some subspace S, then for any vector $\mathbf{z}$, $P\mathbf{z}$ yields the projection of $\mathbf{z}$ onto S.

As a concrete example, suppose $A \in \mathbb{R}^{m\times n}$ has full row rank m. Then $\mathrm{Row}(A) \subseteq \mathbb{R}^n$ has dimension m. The matrix

$P_{\mathrm{Row}(A)} := A^\top(AA^\top)^{-1}A$

is an $n \times n$ matrix that projects any vector in $\mathbb{R}^n$ onto the row space of A. We can verify that it is a projection:

$P_{\mathrm{Row}(A)}^2 = A^\top(AA^\top)^{-1}AA^\top(AA^\top)^{-1}A = A^\top(AA^\top)^{-1}A = P_{\mathrm{Row}(A)}$,

and symmetric:

$P_{\mathrm{Row}(A)}^\top = A^\top((AA^\top)^{-1})^\top A = A^\top(AA^\top)^{-1}A = P_{\mathrm{Row}(A)}$,

since $AA^\top$ is symmetric and invertible. Thus $P_{\mathrm{Row}(A)}$ is an orthogonal projection matrix onto $\mathrm{Row}(A)$.

Similarly, if A has full column rank n, then $AA^\top$ is $m \times m$ invertible and

$P_{\mathrm{Col}(A)} := A(AA^\top)^{-1}A^\top$

is an $m \times m$ projection matrix onto $\mathrm{Col}(A)$.

3 Moore-Penrose Pseudoinverse

When a matrix $X \in \mathbb{R}^{m\times n}$ is not square or not invertible, we cannot talk about the usual inverse X^{-1}. However, we can often define a generalized inverse called the **Moore-Penrose pseudoinverse**, denoted X^+. The pseudoinverse plays a central role in solving least-squares and underdetermined systems.

3.1 Definition and Characterization

The Moore-Penrose pseudoinverse X^+ of a matrix X is defined as the unique matrix (of size $n \times m$) that satisfies the following four properties:

(i) $XX^+X = X$.

(ii) $X^+XX^+ = X^+$.

(iii) $(XX^+)^\top = XX^+$.

(iv) $(X^+X)^\top = X^+X$.

One can show that such a matrix X^+ always exists (for any real matrix X) and is unique. Verifying these four properties ensures you have the pseudoinverse. In many contexts, X^+ behaves like an inverse of X on the appropriate subspaces. In particular, properties (iii) and (iv) imply that $P := X^+X$ and $P' := XX^+$ are orthogonal projection matrices (symmetric idempotent matrices). Specifically:

- $P = X^+X$ is an $n \times n$ projection onto the row space of X (a subspace of $\mathbb{R}^n$).
- $P' = XX^+$ is an $m \times m$ projection onto the column space of X (a subspace of $\mathbb{R}^m$).

Property (i) says X^+ is a *right-inverse* of X on the column space (since $XX^+X = X$ means XX^+ acts like identity on the range of X). Property (ii) says X^+ is a *left-inverse* on the row space (since $X^+XX^+ = X^+$ means X^+X acts like identity on the range of X^+, which is the row space of X).

3.2 Pseudoinverse in the Full Row Rank Case

The above abstract definition can be made concrete in the case we care about: when X has full *row* rank m (with $m \leq n$). In this scenario, $XX^\top$ is an invertible $m \times m$ matrix, whereas $X^\top X$ is singular (since $\text{rank}(X^\top X) = \text{rank}(X) = m < n$).

Proposition 1. *If $X \in \mathbb{R}^{m\times n}$ has full row rank m, then one representation of the pseudoinverse is:*

$$X^+ = X^\top (XX^\top)^{-1}.$$

In other words, X^+ can be written explicitly as the $n \times m$ matrix $X^\top (XX^\top)^{-1}$.

Proof. We need to verify that $X^\top (XX^\top)^{-1}$ satisfies the four defining properties of the pseudoinverse.

(i) $XX^+X = X(X^\top (XX^\top)^{-1})X = (XX^\top)(XX^\top)^{-1}X = I_m X = X$.

(ii) $X^+XX^+ = X^\top (XX^\top)^{-1}XX^\top (XX^\top)^{-1} = X^\top (XX^\top)^{-1}$, since $XX^\top (XX^\top)^{-1} = I_m$. But $X^\top (XX^\top)^{-1}$ is exactly X^+, so property (ii) holds.

(iii) $(XX^+)^\top = (XX^\top (XX^\top)^{-1})^\top = ((XX^\top)(XX^\top)^{-1})^\top = I_m^\top = I_m$. On the other hand, $XX^+ = XX^\top (XX^\top)^{-1} = I_m$. So indeed $(XX^+)^\top = XX^+$.

(iv) $(X^+X)^\top = (X^\top (XX^\top)^{-1}X)^\top = X^\top (XX^\top)^{-1}X$ (since the whole product is a scalar symmetric matrix of size $n \times n$). Thus $(X^+X)^\top = X^+X$.

All four properties are satisfied, so $X^\top (XX^\top)^{-1}$ is the pseudoinverse of X. Uniqueness of the pseudoinverse guarantees that this is the X^+. □

For intuition, note that $X^+ = X^\top (XX^\top)^{-1}$ is precisely the matrix that we identified earlier as the projection-style inverse in the full row rank case. It satisfies $XX^+ = I_m$, so X^+ acts as a right-inverse of X on $\mathbb{R}^m$ (the output space). However, X^+X is not the $n \times n$ identity, but rather the projection onto $\text{Row}(X)$. In fact, using our notation from before, we have $X^+X = P_{\text{Row}(X)}$.

3.3 Least Squares and Minimum-Norm Solutions

The pseudoinverse provides the solution to both least-squares problems and underdetermined linear systems:

- **Least-squares:** For any X and y, the vector X^+y is a solution to the least-squares problem $\min_w \|Xw - y\|_2^2$. If X has full column rank, this reduces to the usual normal equation solution $(X^\top X)^{-1}X^\top y$. If X does not have full column rank (including the overparameterized case), there are infinitely many minimizers of $\|Xw - y\|^2$ (assuming the minimum is zero or the system is consistent). In that case, X^+y gives the one with smallest norm (as we discuss next).

- **Underdetermined systems:** If $Xw = y$ has at least one solution, then $w = X^+ y$ is a solution. Moreover, it is the unique solution with minimum ℓ_2 norm[4]. If $Xw = y$ has no solution (inconsistent system), then $X^+ y$ gives the minimum-norm solution to the least-squares approximation (it yields the w that minimizes $\|Xw - y\|_2$).

We now prove the important fact that $X^+ y$ indeed yields the minimum-norm solution when $Xw = y$ has many solutions.

Theorem 2 (Minimum-Norm Solution via Pseudoinverse)**.** *Suppose $X \in \mathbb{R}^{m \times n}$ has full row rank m, and consider the (underdetermined) linear system $Xw = y$ where $y \in \mathbb{R}^m$ lies in the column space of X. Then among all solutions w to $Xw = y$, the vector*

$w^* := X^+ y = X^\top (XX^\top)^{-1} y$

has the smallest Euclidean norm. Moreover, this w^* is the unique *minimum-norm solution.*

Proof. First, note that $Xw^* = X(X^+ y) = (XX^+)y = I_m y = y$, so w^* is indeed a solution of $Xw = y$. Now let w be *any* solution of $Xw = y$. We will show that $\|w\| \geq \|w^*\|$, with equality if and only if $w = w^*$.

Because both w and w^* are solutions, we have $Xw = Xw^* = y$. Consider the difference $u := w - w^*$. Then $Xu = Xw - Xw^* = y - y = 0$, so $u \in \mathrm{Null}(X)$. That is, the difference between any solution w and the pseudoinverse solution w^* lies in the nullspace of X. Now recall that $w^* = X^+ y = X^\top (XX^\top)^{-1} y$. Since $y \in \mathrm{Col}(X)$, there exists at least one solution; in fact $\mathrm{Col}(X) = \mathbb{R}^m$ (full row rank means columns span $\mathbb{R}^m$), so such a solution exists for every y. The particular w^* we have is in $\mathrm{Row}(X)$, because $w^* = X^\top(\cdot)$ is explicitly in the row space of X. Thus $w^* \in \mathrm{Row}(X)$.

Now $w^* \in \mathrm{Row}(X)$ and $u \in \mathrm{Null}(X)$. By the Fundamental Theorem of Linear Algebra, $\mathrm{Row}(X) \perp \mathrm{Null}(X)$. Therefore w^* is orthogonal to u:

$w^* \cdot u = 0.$

We can now compare the norms:

$\|w\|^2 = \|w^* + u\|^2 = \|w^*\|^2 + 2\, w^* \cdot u + \|u\|^2 = \|w^*\|^2 + \|u\|^2 \geq \|w^*\|^2,$

since $\|u\|^2 \geq 0$. Thus $\|w\| \geq \|w^*\|$. Furthermore, equality holds (i.e. $\|w\| = \|w^*\|$) if and only if $\|u\|^2 = 0$, which means $u = 0$, i.e. $w = w^*$. This shows w^* is the unique solution of minimal norm. □

The above theorem is a key reason the pseudoinverse is so useful: it picks out the "most gentle" solution vector (smallest length) among all those that fit the data perfectly. This minimum-norm property is a form of implicit regularization: although we did not explicitly constrain w to be small, the algebraic procedure of solving $Xw = y$ via X^+ automatically yields the smallest-norm solution[5]. We will see next that gradient-based optimization finds this same solution.

[4] https://en.wikipedia.org/wiki/Moore-Penrose_inverse#Properties

[5] https://en.wikipedia.org/wiki/Moore-Penrose_inverse#Properties

4 Gradient Flow and Linear Regression Setup

We now turn to the optimization dynamics of interest: gradient flow on the squared error objective. We set up the problem of linear regression in the overparameterized regime and introduce gradient flow as an infinite-infinitesimal step size limit of gradient descent.

4.1 Overparameterized Linear Regression

We consider a linear regression problem with m data points and n parameters, in the case $n > m$. Let $X \in \mathbb{R}^{m \times n}$ be the design matrix and $y \in \mathbb{R}^m$ the target vector, exactly as in the introduction. We assume:

1. X has full row rank m. This implies $m \leq n$ and $\operatorname{rank}(X) = m$. Thus the system $Xw = y$ is potentially underdetermined (if $n > m$) but has at least one solution for any y (indeed $\operatorname{Col}(X) = \mathbb{R}^m$ if rank m).

2. The system is *consistent*, meaning there exists at least one solution w such that $Xw = y$. Equivalently, y lies in the column space of X. (This is automatically true since $\operatorname{Col}(X) = \mathbb{R}^m$ by full row rank; we mention it for clarity.)

In this overparameterized setting, there are infinitely many solutions w satisfying $Xw = y$; they form an affine subspace $w^* + \operatorname{Null}(X)$ for any particular solution w^*.

Our goal is to analyze which of these solutions is obtained by *gradient flow* on the squared loss.

4.2 Gradient Flow Dynamics

We define the loss function

$$L(w) = \frac{1}{2}\|Xw - y\|_2^2.$$

(We include the 1/2 factor for mathematical convenience; it does not change the location of minima or the gradient's zeros, only its scale.) Explicitly,

$$L(w) = \frac{1}{2}\sum_{i=1}^{m}(x_i^\top w - y_i)^2.$$

The gradient of $L(w)$ with respect to w is computed using the chain rule:
$\nabla_w L(w) = X^\top (Xw - y)$.
This n-dimensional vector is the derivative of the loss in each parameter direction.
Gradient descent is the discrete-time iterative algorithm:
$w_{k+1} = w_k - \eta \nabla_w L(w_k)$,
for some step size (learning rate) $\eta > 0$. Gradient **flow** is the continuous-time analogue, described by the differential equation (an initial value problem):
$\frac{d}{dt}w(t) = -\nabla_w L(w(t)), \quad w(0) = w_{\text{init}}$.
In our case, this becomes the system of linear ODEs:
$\dot{w}(t) = -X^\top (Xw(t) - y), \quad w(0) = w_{\text{init}}$.

Here $\dot{w}(t)$ denotes the time derivative of $w(t)$.

The gradient flow can be seen as what happens when we take $\eta \to 0$ in gradient descent and simultaneously send the number of steps to infinity so that $\eta k = t$ stays finite. In practice, gradient flow captures the trajectory of gradient descent when the steps are very small. Importantly, gradient flow and gradient descent (with small enough η) will converge to the same limit if they converge, because gradient flow is essentially a smooth interpolation of the discrete updates.

In our analysis, we will focus on the gradient flow dynamics because they are easier to study (being defined by a nice ODE). We will also make a key assumption on the initial condition:

$w(0) = 0.$

That is, we start the parameters at the zero vector. This is a common and "neutral" initialization for linear models (it does not favor any direction in parameter space). It will also be crucial for the implicit bias result: starting at 0 means the initial parameter vector has no component in any particular direction, including the nullspace.

5 Main Theorem

We can now state the central result of these notes.

Theorem 3 (Gradient Flow Converges to Minimum-Norm Solution)**.** *Consider the over-parameterized linear regression setup with loss $L(w) = \frac{1}{2}\|Xw - y\|^2$ as above, and let $w(t)$ follow the gradient flow $\dot{w}(t) = -X^\top(Xw(t) - y)$ with initial condition $w(0) = 0$. Then as $t \to \infty$, $w(t)$ converges to $w^* = X^+y$, the unique minimum-norm solution of $Xw = y$. In other words,*

$$\lim_{t\to\infty} w(t) = X^+y,$$

and this limit w^* satisfies $Xw^* = y$ and $\|w^*\| = \min\{\|w\| : Xw = y\}$.

A few remarks before we proceed to the proof:

- The theorem implies that gradient flow (and by extension, gradient descent for sufficiently small step sizes) will find the pseudoinverse solution X^+y among all possible solutions. This is precisely the statement of implicit regularization for linear regression: even without an explicit norm penalty, the algorithm biases towards the minimum-norm interpolant[6].

- The assumption $w(0) = 0$ is important. In fact, if we started with some nonzero $w(0)$ that had a component in the nullspace of X, that component would *never* be corrected by the gradient flow, because the gradient $X^\top(Xw - y)$ is always orthogonal to the nullspace of X. In such a case, the limit would be X^+y plus the persistent nullspace component of the initial $w(0)$[7]. By starting at 0 (which has no nullspace component), we ensure the trajectory stays within the row space of X and thus converges to the pure minimum-norm solution (which lies in the row space).

[6] `https://math.stackexchange.com`

[7] `https://math.stackexchange.com`

- The condition that X has full row rank can be relaxed slightly. If X does not have full row rank (so the system might not be consistent or the column space is a strict subset of $\mathbb{R}^m$), gradient flow will still converge to X^+y, which in that case is the least-squares minimizer (not achieving zero training error unless y has no component in the orthogonal complement of $\mathrm{Col}(X)$). For simplicity, we assume full row rank so that y is exactly attainable.

6 Proof of the Main Theorem

We now give a complete proof of Theorem 3. The proof will involve examining the structure of the gradient flow ODE and leveraging the linear algebra results from earlier sections. We will break the reasoning into a few steps and lemmas for clarity.

Step 1: Decomposition into Row Space and Nullspace

Recall that $\mathbb{R}^n = \mathrm{Row}(X) \oplus \mathrm{Null}(X)$ (direct sum decomposition). Since $w(0) = 0$, we can express $w(t)$ for any t uniquely as $w_{\text{row}}(t) + w_{\text{null}}(t)$, where $w_{\text{row}}(t) \in \mathrm{Row}(X)$ and $w_{\text{null}}(t) \in \mathrm{Null}(X)$. We will show that actually $w_{\text{null}}(t)$ remains zero for all time if it starts zero.

Lemma 1 (Nullspace Component Remains Constant)**.** *Under the gradient flow $\dot{w}(t) = -X^\top(Xw(t) - y)$, the component of $w(t)$ in* $\mathrm{Null}(X)$ *does not change over time. In fact, $\frac{d}{dt}w_{\text{null}}(t) = 0$ for all t.*

Proof. Take any vector $v \in \mathrm{Null}(X)$ (so $Xv = 0$). Consider the dot product $v \cdot \dot{w}(t)$:

$v \cdot \dot{w}(t) = v \cdot [-X^\top(Xw(t) - y)] = -(Xv) \cdot (Xw(t) - y) = -0 \cdot (Xw(t) - y) = 0.$

Thus $v \cdot \dot{w}(t) = 0$ for all v in the nullspace of X. But $w_{\text{null}}(t)$ is, by definition, the projection of $w(t)$ onto $\mathrm{Null}(X)$. The time derivative of that projection is simply $\frac{d}{dt}w_{\text{null}}(t) = \Pi_{\mathrm{Null}(X)}(\dot{w}(t))$. However, $\dot{w}(t)$ has zero inner product with every vector in $\mathrm{Null}(X)$, which means $\dot{w}(t)$ is orthogonal to $\mathrm{Null}(X)$ at all times. Therefore, the projection of $\dot{w}(t)$ onto $\mathrm{Null}(X)$ is zero. In formula:

$\frac{d}{dt}w_{\text{null}}(t) = \Pi_{\mathrm{Null}(X)}(\dot{w}(t)) = 0.$

This shows that $w_{\text{null}}(t)$ is constant in time. Given that $w_{\text{null}}(0) = 0$ (because the initial $w(0)$ has no nullspace component), we conclude $w_{\text{null}}(t) = 0$ for all t. □

This lemma formalizes the intuition that gradient flow never produces movement in directions that do not affect the loss. Since any movement in the nullspace of X does not change Xw (and hence does not change the loss $L(w)$), the gradient in those directions is zero, so any nullspace component of w remains as whatever it was initially. In our case, that means $w(t)$ stays entirely in $\mathrm{Row}(X)$ for all time (because we started with no nullspace component).

Step 2: Dynamics within the Row Space

Due to Lemma 1, we can restrict our attention to the row space. When $w(t) \in \mathrm{Row}(X)$, we can parametrize $w(t)$ in terms of an m-dimensional vector (since $\dim(\mathrm{Row}(X)) = m$).

Specifically, because the columns of $X^\top$ span the row space, there exists some vector $u(t) \in \mathbb{R}^m$ such that:

$w(t) = X^\top u(t)$.

(If X has full row rank, $X^\top$ has linearly independent columns, so this representation $u(t)$ is unique.)

Our plan will be to derive a differential equation for $u(t)$ and solve it. First, let's express $Xw(t) - y$ in terms of $u(t)$:

$Xw(t) - y = X(X^\top u(t)) - y = (XX^\top)u(t) - y$.

Using this, the gradient flow equation $\dot{w}(t) = -X^\top(Xw(t) - y)$ becomes:

$$\begin{aligned}\dot{w}(t) &= -X^\top[(XX^\top)u(t) - y] \\ &= -X^\top(XX^\top)u(t) + X^\top y.\end{aligned}$$

But $w(t) = X^\top u(t)$, so differentiating both sides gives:

$\dot{w}(t) = X^\top \dot{u}(t)$.

Equating the two expressions for $\dot{w}(t)$, we get:

$X^\top \dot{u}(t) = -X^\top(XX^\top)u(t) + X^\top y$.

We can cancel $X^\top$ on the left side (more rigorously, since $X^\top$ is one-to-one on $\mathbb{R}^m$, we can premultiply both sides by $(X^\top)^+ = (XX^\top)^{-1}X$ to get equivalently):

$\dot{u}(t) = -(XX^\top)u(t) + y$.

This is now a linear ODE in m-dimensional space:

$$\dot{u}(t) + (XX^\top)u(t) = y. \tag{$*$}$$

The matrix $XX^\top$ is an $m \times m$ symmetric positive-definite matrix (since X has rank m). Equation $(*)$ is a first-order linear inhomogeneous ODE with a constant coefficient matrix $XX^\top$ and constant "forcing" term y.

We can solve this ODE using standard methods. The homogeneous part $\dot{u} + (XX^\top)u = 0$ has general solution $u_{\text{hom}}(t) = e^{-(XX^\top)t}c$ for some constant vector $c \in \mathbb{R}^m$ (here $e^{-(XX^\top)t}$ is the matrix exponential of $-(XX^\top)t$). To find a particular solution to the full equation, note that since the forcing term y is constant in time, we can look for a constant solution $u_{\text{part}}(t) = u^*$ (a constant vector). Plugging u^* into $(*)$ and setting $\dot{u} = 0$, we get:

$0 + (XX^\top)u^* = y$.

Since $XX^\top$ is invertible, this gives $u^* = (XX^\top)^{-1}y$.

Thus the general solution to $(*)$ is:

$u(t) = e^{-(XX^\top)t}[-(XX^\top)^{-1}y] + (XX^\top)^{-1}y = (I_m - e^{-(XX^\top)t})(XX^\top)^{-1}y$.

At $t = 0$, we have $w(0) = 0$, hence $0 = w(0) = X^\top u(0)$, which implies $X^\top u(0) = 0$. Because X has full row rank, $X^\top$ is one-to-one, so $u(0)$ must be the zero vector in $\mathbb{R}^m$. Therefore $u(0) = 0$. Using the general solution for $u(t)$:

$u(0) = -(XX^\top)^{-1}y + (XX^\top)^{-1}y = 0$,

so $c = -(XX^\top)^{-1}y$.

Now we have the specific solution for $u(t)$:

$u(t) = e^{-(XX^\top)t}[-(XX^\top)^{-1}y] + (XX^\top)^{-1}y = (I_m - e^{-(XX^\top)t})(XX^\top)^{-1}y$.

Finally, recall that $w(t) = X^\top u(t)$. Thus:

$$w(t) = X^\top \Big(I_m - e^{-(XX^\top)t}\Big)(XX^\top)^{-1}y. \tag{$\dagger$}$$

At $t = 0$, $w(0) = X^\top(I_m - I_m)(XX^\top)^{-1}y = 0$, as expected. As $t \to \infty$, note that $e^{-(XX^\top)t} \to 0$ (the matrix exponential decays to the zero matrix because $XX^\top$ is positive-definite with real eigenvalues $\lambda_i > 0$, so $e^{-\lambda_i t} \to 0$ for each eigenvalue). Thus $I_m - e^{-(XX^\top)t} \to I_m$. Therefore

$\lim_{t\to\infty} w(t) = X^\top(I_m)(XX^\top)^{-1}y = X^\top(XX^\top)^{-1}y.$

But $X^\top(XX^\top)^{-1}y$ is exactly the pseudoinverse solution X^+y. So indeed $\lim_{t\to\infty} w(t) = X^+y$.

We have thus shown that $w(t)$ converges to X^+y as $t \to \infty$. This proves that gradient flow reaches the minimum-norm solution.

To be completely thorough, we should verify that the limit $w^* = X^\top(XX^\top)^{-1}y$ indeed satisfies the normal equations (which guarantee it is a stationary point of the loss) and that it fits the data:

- $Xw^* = XX^\top(XX^\top)^{-1}y = y$, so w^* achieves zero training error (it interpolates the data).

- $X^\top(Xw^* - y) = X^\top(y - y) = 0$, so the gradient $\nabla L(w^*) = 0$. Thus w^* is a critical point of the loss. In fact, w^* is the global minimizer of the loss $L(w)$ on the affine set $\{w : Xw = y\}$, but since all points on that affine set have $L(w) = 0$, any of them is a global minimizer of L. The uniqueness of w^* comes not from the objective value but from the minimal norm consideration, which we already addressed in Theorem 2.

This completes the proof of Theorem 3.

7 Intuition and Interpretation

The above proof, while algebraically precise, can be understood in simpler terms. At its heart, the reason gradient flow finds the minimum-norm solution is because it only moves in directions that reduce the training error, and those directions turn out to be exactly the directions within the row space of X. Any component of w in the nullspace of X does not affect the output Xw at all, so gradient flow has no incentive to move in those directions. In fact, as we showed, if you start with $w(0) = 0$ (which has no nullspace component), you will never pick up a nullspace component as the dynamics evolve. This means the solution $w(t)$ always stays in the row space. Among all solutions of $Xw = y$, the one that lies in the row space is exactly the minimum-norm solution (recall Theorem 2). Thus, gradient flow implicitly imposes the condition that w stays in $\mathrm{Row}(X)$, which is why it ends up at X^+y and not any other solution.

Another perspective is to consider an explicit ℓ_2 regularization and then remove it. If we add a small weight-decay term $\frac{\lambda}{2}\|w\|^2$ to the loss (where $\lambda > 0$), the minimizer becomes

$w_\lambda = \arg\min_w \left\{\frac{1}{2}\|Xw - y\|^2 + \frac{\lambda}{2}\|w\|^2\right\}.$

Setting the gradient to zero gives the so-called *ridge regression* normal equations:

$X^\top(Xw_\lambda - y) + \lambda w_\lambda = 0,$

which leads to

$w_\lambda = (X^\top X + \lambda I_n)^{-1}X^\top y,$

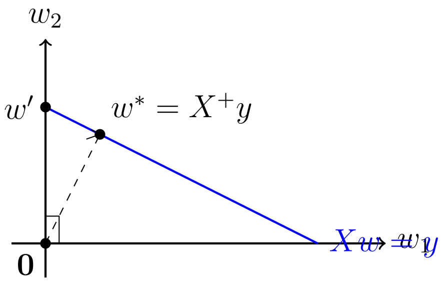

Figure 1: Geometric illustration in $\mathbb{R}^2$. The line represents the affine subspace of solutions $\{w : Xw = y\}$ for an underdetermined system. Gradient flow starting at $\mathbf{0}$ will move along the perpendicular direction to reach the solution w^* directly, rather than sliding along the line to some other w'. Thus w^*, the orthogonal projection of $\mathbf{0}$ onto the solution set, is the one with minimum norm.

assuming n is not too large or we interpret the inverse in a generalized sense. For $\lambda > 0$, $X^\top X + \lambda I$ is invertible (even if $X^\top X$ is singular) because λI makes it strictly positive-definite. As $\lambda \to 0^+$, one can show that $w_\lambda \to X^+ y$[8]. This aligns with our gradient flow result: the minimum-norm interpolating solution is what you get by taking ridge regression and letting the regularization weight go to zero. Gradient descent implicitly performs this limit: it finds the same w^* without you ever having to specify λ explicitly.

In practical terms, this implicit regularization explains why even very complex models (with many more parameters than data points) can generalize well: the training algorithm itself selects a "simple" solution (in linear regression, the one of smallest norm) out of the many that fit the training data perfectly. In more advanced settings (like neural networks), analyzing implicit bias is much harder, but the linear case we proved here is one of the first and most important examples illustrating the concept.

8 Exercises

1. **Verification of Pseudoinverse Properties.** Let $X = \begin{pmatrix} 1 & 2 & 0 \\ 0 & 2 & 2 \end{pmatrix}$, which is a 2×3 full row rank matrix. Compute its pseudoinverse X^+ using the formula from Proposition 1. Then verify directly (by matrix multiplication) that X^+ satisfies the four Moore-Penrose conditions (i)–(iv) from Section 3.

2. **Minimum-Norm Solution in a Simple System.** Consider a linear system with one equation and two unknowns: $x_1 + 2x_2 = 4$. Describe the set of all solutions (x_1, x_2) in $\mathbb{R}^2$. Find the solution with minimum Euclidean norm. (Hint: You can do this by geometry, or by using the pseudoinverse.) Now suppose we run gradient descent on $L(w) = \frac{1}{2}(x_1 + 2x_2 - 4)^2$ starting from $(0, 0)$. If we use a small step size, which solution does it approach?

3. **Influence of Initialization.** Let $w(0) = w_0$ be an arbitrary initial vector, and $w_0 =$

[8] `https://math.stackexchange.com`

$w_0^{\parallel} + w_0^{\perp}$ be its decomposition into $\mathrm{Row}(X)$ (the part $w_0^{\parallel}$) and $\mathrm{Null}(X)$ (the part $w_0^{\perp}$). Without doing any new calculations, use our results to argue what $\lim_{t\to\infty} w(t)$ will be for gradient flow started at this $w(0)$. In particular, how does the limit depend on $w_0^{\perp}$? What happens if $w_0^{\perp} \neq 0$? (You may assume X has full row rank as usual.)