Implicit Bias of Gradient Descent Toward Minimum-Norm Solution

Hossein Mobahi
Google Research
Mountain View, CA, USA
hmobahi@google.com

September 23, 2021

1 Introduction

You might have heard about implicit regularization of SGD; meaning preferences toward certain solutions that are not explicitly stated in the objective function. One of them is that SGD prefers solutions with small norm. In fact, such minimum norm bias is not a product of stochasticity of SGD, but the gradient descent itself. Where does this bias come from?

Here I analyze a simple scenario: linear regression with squared loss. We know gradient descent has min-norm bias in this case and we can simply prove that. I first derive the min-norm solution for this problem, and then show gradient descent converges to that.

2 Minimum-Norm Solution

Consider training data $(\boldsymbol{x}_k, y_k)$ for $k = 1, \ldots, n$, where each $\boldsymbol{x}_k \in \mathbb{R}^d$ and $y_k \in \mathbb{R}$. Suppose we are in over-parameterized regime, i.e. $d > n$. The goal is to find a $\boldsymbol{w}$ that interpolates the data,

$$\boldsymbol{x}_k^T \boldsymbol{w} = y_k \quad \text{for} \quad k = 1, \ldots n \,, \tag{1}$$

or equivalently in matrix form,

$$\boldsymbol{X}^T \boldsymbol{w} = \boldsymbol{y} \,, \tag{2}$$

where $\boldsymbol{X}_{d\times n} = \big[\boldsymbol{x}_1 \mid \ldots \mid \boldsymbol{x}_n\big]$ and $\boldsymbol{y}_{n\times 1} = \big[y_1 \mid \ldots \mid y_n\big]$. Since $d > n$, there are possibly infinitely many $\boldsymbol{w}$'s that satisfy the above. To keep the problem well-posed, we seek the $\boldsymbol{w}$ that has the smallest norm,

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{w}\|^2 \quad \text{s.t.} \quad \boldsymbol{X}^T \boldsymbol{w} = \boldsymbol{y} \,, \tag{3}$$

Using the method of Lagrange multiplier, it is straightforward[1] to show that the solution has the form,

$$\boldsymbol{w}^* \;=\; \boldsymbol{X}(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{y} \;. \tag{11}$$

3 Gradient Descent

Suppose we wish to find the $\boldsymbol{w}$ satisfying the $\boldsymbol{X}^T \boldsymbol{w} = \boldsymbol{y}$ using gradient descent with step size $\eta > 0$ by minimizing the squared loss J,

$$J(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{X}^T \boldsymbol{w} - \boldsymbol{y}\|^2 \,. \tag{12}$$

We assume the data matrix $\boldsymbol{X}$ is full-rank, i.e. $\text{rank}(\boldsymbol{X}) = n$. Since $d > n$, there are infinitely many $\boldsymbol{w}$'s that can interpolate the data and achieve zero loss. Note that we are ***not explicitly enforcing any regularization*** here to prefer one solution versus the other. We now show that the solution of gradient descent converges to the min-norm solution $\boldsymbol{w}^*$ in (11).

[1] Denote the Lagrange multiplier vector by $\boldsymbol{\lambda} \in \mathbb{R}^n$. The Lagrangian can be expressed as,

$$L(\boldsymbol{w}, \lambda) = \frac{1}{2}\|\boldsymbol{w}\|^2 + \langle \boldsymbol{\lambda}, \boldsymbol{X}^T \boldsymbol{w} - \boldsymbol{y} \rangle \,. \tag{4}$$

Since,

$$\nabla_{\boldsymbol{w}} L(\boldsymbol{w}, \lambda) \;=\; \boldsymbol{w} + \boldsymbol{X}\boldsymbol{\lambda} \tag{5}$$
$$\nabla_{\boldsymbol{\lambda}} L(\boldsymbol{w}, \lambda) \;=\; \boldsymbol{X}^T \boldsymbol{w} - \boldsymbol{y} \,. \tag{6}$$

By zero crossing the above we can obtain the optimality condition,

$$\boldsymbol{w} = -\boldsymbol{X}\boldsymbol{\lambda} \quad , \quad \boldsymbol{X}^T \boldsymbol{w} = \boldsymbol{y} \;. \tag{7}$$

Replacing $\boldsymbol{w}$ implies,

$$-\boldsymbol{X}^T \boldsymbol{X}\, \boldsymbol{\lambda} = \boldsymbol{y} \;. \tag{8}$$

Since the $n \times n$ matrix $\boldsymbol{X}^T \boldsymbol{X}$ is full-rank (as $\boldsymbol{X}$ has rank n and $d > n$), its inverse is ***well-defined*** and thus,

$$\boldsymbol{\lambda} = -(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{y} \;. \tag{9}$$

Replacing $\boldsymbol{\lambda}$ yields the optimal $\boldsymbol{w}$,

$$\boldsymbol{w} = -\boldsymbol{X}\boldsymbol{\lambda} = \boldsymbol{X}(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{y} \;. \tag{10}$$

The gradient descent updates have the form,

$$\begin{aligned}
\boldsymbol{w}_{t+1} &= \boldsymbol{w}_t - \eta \nabla J(\boldsymbol{w}_t) && (13)\\
&= \boldsymbol{w}_t - \eta \boldsymbol{X}\Big(\boldsymbol{X}^T \boldsymbol{w}_t - \boldsymbol{y}\Big) && (14)\\
&= \Big(\boldsymbol{I} - \eta \boldsymbol{X}\boldsymbol{X}^T\Big)\boldsymbol{w}_t + \eta \boldsymbol{X}\boldsymbol{y}\,. && (15)
\end{aligned}$$

From here it is easy to resolve the recurrence and obtain,

$$\boldsymbol{w}_t \;=\; (\boldsymbol{I} - \eta \boldsymbol{X}\boldsymbol{X}^T)^t \boldsymbol{w}_0 + \eta\Big(\sum_{i=0}^{t-1} (\boldsymbol{I} - \eta \boldsymbol{X}\boldsymbol{X}^T)^i\Big)\boldsymbol{X}\boldsymbol{y}\,. \tag{16}$$

Since $d > n$ and $\text{rank}(\boldsymbol{X}) = n$ (full-rank assumption), $(\boldsymbol{X}^T\boldsymbol{X})^{-\frac{1}{2}}$ is well-defined. This enables us to have the identity[2],

$$(\boldsymbol{I}_{d\times d} - \eta \boldsymbol{X}\boldsymbol{X}^T)^i \boldsymbol{X} \;=\; \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-\frac{1}{2}}(\boldsymbol{I}_{n\times n} - \eta \boldsymbol{X}^T\boldsymbol{X})^i(\boldsymbol{X}^T\boldsymbol{X})^{\frac{1}{2}}\,. \tag{19}$$

Applying this identity to the weight update yields,

$$\boldsymbol{w}_t \;=\; (\boldsymbol{I} - \eta \boldsymbol{X}\boldsymbol{X}^T)^t \boldsymbol{w}_0 + \eta \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-\frac{1}{2}}\Big(\sum_{i=0}^{t-1} (\boldsymbol{I} - \eta \boldsymbol{X}^T\boldsymbol{X})^i\Big)(\boldsymbol{X}^T\boldsymbol{X})^{\frac{1}{2}}\boldsymbol{y}\,. \tag{20}$$

The series $\sum_{i=0}^{\infty}(\boldsymbol{I} - \eta \boldsymbol{X}^T\boldsymbol{X})^i$ converges when $\|\boldsymbol{I} - \eta \boldsymbol{X}^T\boldsymbol{X}\| < 1$, which is equivalent to $|1 - \eta\lambda| < 1$ where λ is an eigenvalue of the matrix $\boldsymbol{X}^T\boldsymbol{X}$. This condition can be satisfied by choosing learning rate small enough: $\eta < \frac{2}{\lambda_{\max}}$ when $\lambda_{\min} \neq 0$ (the latter requirement is the reason we switched from the rank-deficient $d \times d$ matrix $\boldsymbol{X}\boldsymbol{X}^T$ to the full rank $n \times n$ matrix $\boldsymbol{X}^T\boldsymbol{X}$).

The series $\sum_{i=0}^{\infty} \boldsymbol{A}^i$ is called ***Neumann series***, and when convergent ($\|\boldsymbol{A}\| < 1$), it converges to $(\boldsymbol{I} - \boldsymbol{A})^{-1}$. Thus, under the initialization assumption $\boldsymbol{w}_0 = \boldsymbol{0}$, it follows that, when $\eta < \frac{2}{\lambda_{\max}}$,

$$\begin{aligned}
\boldsymbol{w}_\infty &= (\boldsymbol{I} - \eta \boldsymbol{X}\boldsymbol{X}^T)^\infty \boldsymbol{w}_0 + \eta \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-\frac{1}{2}}(\eta \boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{X})^{\frac{1}{2}}\boldsymbol{y} && (21)\\
&= \eta \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-\frac{1}{2}}(\eta \boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{X})^{\frac{1}{2}}\boldsymbol{y} && (22)\\
&= \eta \boldsymbol{X}\frac{1}{\eta}(\boldsymbol{X}^T\boldsymbol{X})^{-\frac{1}{2}-1+\frac{1}{2}}\boldsymbol{y} && (23)\\
&= \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{y}\,, && (24)
\end{aligned}$$

where in (21) we use the fact that $\lim_{i\to\infty} \|(\boldsymbol{I} - \eta \boldsymbol{X}\boldsymbol{X}^T)^i \boldsymbol{w}_0\| = 0$ because $\|\boldsymbol{I} - \eta \boldsymbol{X}^T\boldsymbol{X}\| \leq 1$ is bounded and $\boldsymbol{w}_0 = \boldsymbol{0}$. We just proved the following theorem.

Theorem 1 *Consider linear regression with squared loss $\frac{1}{2}\|\boldsymbol{X}^T\boldsymbol{w} - \boldsymbol{y}\|^2$ in the over-parameterized regime $d > n$. Suppose the data matrix $\boldsymbol{X}$ is full-rank, i.e. $\text{rank}(\boldsymbol{X}) = n$. Let $\lambda_{\max}$ be the largest eigenvalue of the matrix $\boldsymbol{X}^T\boldsymbol{X}$. Then for any step satisfying $0 < \eta < \frac{2}{\lambda_{\max}}$ under initialization $\boldsymbol{w}_0 = \boldsymbol{0}$, $\boldsymbol{w}$ converges to the solution of the problem $\min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{w}\|^2$ s.t. $\boldsymbol{X}^T\boldsymbol{w} = \boldsymbol{y}$.*

[2]This identity can be easily proved. Consider singular value decomposition of $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T$. It is straightforward to verify that,

$$\begin{aligned}
(\boldsymbol{I}_{d\times d} - \eta \boldsymbol{X}\boldsymbol{X}^T)^i \boldsymbol{X} &= \boldsymbol{U}(\boldsymbol{I}_{d\times d} - \eta \boldsymbol{S}\boldsymbol{S}^T)^i \boldsymbol{S}\boldsymbol{V}^T && (17)\\
\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-\frac{1}{2}}(\boldsymbol{I}_{n\times n} - \eta \boldsymbol{X}^T\boldsymbol{X})^i(\boldsymbol{X}^T\boldsymbol{X})^{\frac{1}{2}} &= \boldsymbol{U}\boldsymbol{S}(\boldsymbol{I}_{n\times n} - \eta \boldsymbol{S}^T\boldsymbol{S})^i \boldsymbol{V}^T\,. && (18)
\end{aligned}$$

It is easy to verify that $(\boldsymbol{I}_{d\times d} - \eta \boldsymbol{S}\boldsymbol{S}^T)^i \boldsymbol{S} = \boldsymbol{S}(\boldsymbol{I}_{n\times n} - \eta \boldsymbol{S}^T\boldsymbol{S})^i$ (recall that $\boldsymbol{S}_{d\times n}$ consists of a $n \times n$ diagonal block, and is zero elsewhere). This implies that the left hand side of (17) and (18) are identical.