Theoretical Deep Learning Report

Nafissa BENALI & Angel REYERO LOBO
Teacher: Hédi HADIJI

Institut Mathématique d'Orsay(IMO)
Paris-Saclay University

February 13, 2024

Abstract

There is a scarcity of theoretical results on the implicit bias induced by SGD and GD that does not neglect the effect of stochasticity and stepsize. To address this, we present the ideas of Even et al. (2023) in the simple case of 2-layer Diagonal Linear Networks. This case may provide insights into the implicit bias of SGD. In fact, we demonstrate that with small stepsizes, both SGD and GD are close to the Gradient Flow approximation, while with a macroscopic stepsize, the solutions differ. Furthermore, while SGD achieves better performance, GD is unable to recover the true parameters.

Keywords: (Stochastic) Gradient Descent, Edge of Stability, Gradient Flow, Diagonal Linear Networks.

Notations. We denote $\odot$ the element-wise product (Hadamard product). The Bregman divergence of a differentiable convex function $h : \mathbb{R}^d \to \mathbb{R}$ is given by $\mathcal{D}_h(\beta_1, \beta_2) = h(\beta_1) - h(\beta_2) - \langle \nabla h(\beta_2), \beta_1 - \beta_2 \rangle$.

Contents

1 Introduction

1.1 Context

This project primarily relies on the research conducted by Even et al. (2023). It is undertaken as part of the evaluation for the *Theoretical Deep Learning* course of CentraleSupélec in the second year of the Master's program in *Mathematics and Artificial Intelligence* at the Institut Mathématique d'Orsay (IMO) at Paris-Saclay University.

Additionally, we have created a corresponding Github repository containing the code used to generate all the figures in this document. This repository aims to enhance the interpretability of the results.

1.2 Motivations

The *Stochastic Gradient Descent (SGD)* algorithm, although simple, is widely used in the domain of Deep Learning. It is observed that in many contexts, where infinite number of solutions are possible, such as overparameterized linear regression where the set of trainable parameters is larger than the sample size, the SGD does not converge to just any solution. In fact, even without explicit regularization, it is observed that the solutions provided by the optimization problem tend to generalize well.

To study this implicit bias, many approaches have been proposed. Primarily, the *Gradient Flow* approach, which is a continuous-time limit simplification of the problem. However, such approaches neglect the stochasticity and the stepsize of the model. Many experimental results have highlighted the importance of these parameters.

Therefore, the objective of this document is to summarize the theoretical analysis presented by Even et al. (2023) in order to address this gap, specifically in the context of a neural network presented in Section 1.3. The main idea is that the SGD solution generalizes better than the GD because of the flatness of the minima picked by SGD in the Edge of Stability Regime, which is the narrow window of stepsize before the divergence of the algorithm. This fact can be empirically observed in Figure 1.

In Section 1.3, we begin by introducing the setup for our study. Section 2 presents the main theorem regarding the convergence and estimates provided by SGD and GD. Finally, Section 3 interprets the results presented to explain Figure 1.

1.3 Setting

Overparametrised linear regression: We want to compute the output $y \in \mathbb{R}^n$ from the input $X \in (\mathbb{R}^d)^n$, where the input dimension d is much larger than the number of samples n. Then, there are infinite interpolators, i.e., $\beta^\star$ such that $y_i = \langle \beta^\star, x_i \rangle$. We denote

$$\mathcal{S} := \left\{ \beta^\star \in \mathbb{R}^d \text{ st } \langle \beta^\star, x_i \rangle = y_i, \forall i \in [n] \right\}$$

the set of interpolators.

2-Layer linear diagonal network: We consider the parametrisation $\langle \beta, x \rangle = \langle u \odot v, x \rangle$, where $w = (u, v) \in \mathbb{R}^{2d}$ are the weights. It can be seen as a simple neural network with activation function σ the identity: $x \to \langle u \odot v, x \rangle = \langle u, \sigma(\mathrm{diag}(v)x) \rangle$.

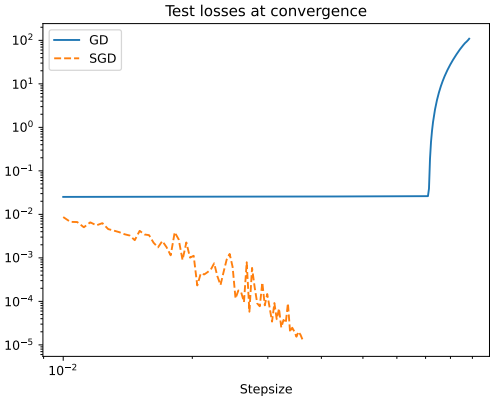

Figure 1: Generalization difference between GD and SDG for DLN. The inputs are Gaussian and the output is $y_i = \langle \beta^\star, x_i \rangle$ for some $s-$sparse vector $\beta^\star$. We take uniform initialisation $\boldsymbol{\alpha} = \alpha \mathbf{1}$ and $(n, d, s, \alpha) = (20, 30, 3, 0.1)$.

Quadratic loss: For a minibatch $\mathcal{B}_k \subset [n]$ of size b, we define the loss function on the batch as

$$F_{\mathcal{B}}(w) := \mathcal{L}_{\mathcal{B}}(u \odot v) = \frac{1}{2b} \sum_{i \in \mathcal{B}} (y_i - \langle u \odot v, x_i \rangle)^2.$$

Mini-batch SGD: We will obtain the final weights by iterating $w_0 = (u_0, v_0)$, $w_{k+1} = w_k - \gamma_k \nabla F_{\mathcal{B}_k}(w_k)$, where γ_k are the stepsizes and the mini-batches $\mathcal{B}_k \subset [n]$ of size b are sampled uniformly and independently. We will take $u_0 = \sqrt{2}\boldsymbol{\alpha}$ and $v_0 = \mathbf{0}$.

2 Implicit bias of (S)GD

In this section, we begin by presenting a result concerning the Gradient Flow approach as outlined by Woodworth et al. (2020). We also provide some insights to interpret the outcome based on the algorithm's initialization. Subsequently, we introduce the main result: the convergence and characterization of the algorithm's outcome as a solution to a regularization problem.

2.1 Gradient Flow

Theorem 2.1 (Gradient Flow on DLN)**.** *the limit $\beta^\star_\alpha$ of the Gradient Flow $\mathrm{d}w_t = -\nabla F(w_t)\mathrm{d}t$ initialised at $(u_0, v_0) = (\sqrt{2}\boldsymbol{\alpha}, \mathbf{0})$ is the solution of the minimal interpolation problem:*

$$\beta^\star_\alpha = \underset{\beta^\star \in \mathcal{S}}{\operatorname{argmin}} \psi_\alpha(\beta^\star) \text{ where } \psi_\alpha(\beta) := \frac{1}{2} \sum_{i=1}^{d} \left(\beta_i \operatorname{arcsinh}\left(\frac{\beta_i}{\alpha_i^2} \right) - \sqrt{\beta_i^2 + \alpha_i^4} + \alpha_i^2 \right)$$

is the hyperbolic entropy function (or hyperentropy).

The generalisation properties mainly rely on two properties of the vector $\boldsymbol{\alpha}$:

Scale of $\boldsymbol{\alpha}$: To see the importance, suppose the uniform initialization $\boldsymbol{\alpha} = \alpha \mathbf{1}$. Then, as α vanishes to 0, $\psi_\alpha \sim \ln(1/\alpha) \|\cdot\|_1$. As usual with the geometry induced by the ℓ_1-norm, this provides sparse recovery properties.

Shape of $\boldsymbol{\alpha}$: In fact, we have that $\psi_\alpha(\beta) \sim \sum_{i=1}^d \ln(1/\alpha_i) |\beta_i|$ as $\boldsymbol{\alpha} \to \mathbf{0}$. Then, if the coordinates do not converge uniformly to 0, this hyperentropy will be a weighted ℓ_1-norm.

2.2 Main result: implicit bias and convergence of the (S)GD

Theorem 2.2 (Implicit bias and convergence)**.** *: Let $(u_k, v_k)_{k\geq 0}$ follow the mini-batch SGD recursion and assume that $(\beta_k)_{k\geq 0} = (u_k \odot v_k)_{k\geq 0}$ converge to some interpolator $\beta^\star_\infty \in \mathcal{S}$. Then,*

$$\beta^\star_\infty = \underset{\beta^\star \in \mathcal{S}}{\operatorname{argmin}} \mathcal{D}_{\psi_{\alpha_\infty}} \left(\beta^\star, \widetilde{\beta}_0\right),$$

with $\widetilde{\beta}_0$ a small perturbation term and $\mathcal{D}_{\psi_{\alpha_\infty}}$ is the Bregman divergence with hyperentropy ψ_{α_∞} of the effective initialization

$$\alpha^2_\infty = \alpha^2 \odot \exp\left(-\sum_{k=0}^{\infty} q\left(\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)\right)\right)$$

and $q(x) = -\frac{1}{2}\ln\left((1-x^2)^2\right)$. Moreover, under the assumption that the stepsize is smaller than a threshold depending on a smoothness parameter, the iterates converge to this interpolator $\beta^\star_\infty$.

The first part of the theorem demonstrates that if the iterates converge to an interpolator, then this interpolator is the solution of a minimization problem that depends on the initialization, the stepsizes, and the trajectory of the iterates. Subsequently, the second part of the theorem establishes convergence. As we will observe in the sequence, the importance of this result lies in the fact that the minimization problem depends in a way that we can understand on the stochasticity and the stepsizes.

First, we use that as proved in Even et al. (2023), under some reasonable assumptions, $\widetilde{\beta}_0$ can be roughly ignored. Then, we observe that

$$\beta^\star_\infty = \underset{\beta^\star \in \mathcal{S}}{\operatorname{argmin}} \mathcal{D}_{\psi_{\alpha_\infty}} \left(\beta^\star, \widetilde{\beta}_0\right) \approx \underset{\beta^\star \in \mathcal{S}}{\operatorname{argmin}} \mathcal{D}_{\psi_{\alpha_\infty}} (\beta^\star, 0) = \underset{\beta^\star \in \mathcal{S}}{\operatorname{argmin}} \psi_{\alpha_\infty}(\beta^\star).$$

This is the solution given by the GF (see Theorem 2.1) where the initialization α is changed by α_∞. Then, we observe that if the stepsize tends to 0, then

$$\exp\left(-\sum_{k=0}^{\infty} q\left(\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)\right)\right) \to \mathbf{1},$$

so that $\alpha_\infty \to \alpha$ and we have the same solution given by GF.

As shown in Even et al. (2023), it is even possible to show some convergence rates using a time varying stochastic mirror descent recursion.

3 Analysis of the result

In this section we analyse the impact of stochasticity and stepsize using Theorem 2.2. To do so, we suppose constant stepsize γ and uniform initialization of $\boldsymbol{\alpha} = \alpha\mathbf{1}$. As seen in the previous section, the deviation from the GF will be mainly controlled by $\sum_{k=0}^{\infty} q\left(\gamma \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)\right)$. In the sequel, we will denote this quantity the Gain_γ.

We first observe that the **scale** of Gain_γ distinguishes GF with (S)GD, and if $\|\text{Gain}_\gamma\|_1 \approx 0$, then they are both similar, as $\alpha_\infty \approx \alpha$.

Similarly as in Section 2.1, we can also study the impact of the **shape** of Gain_γ. Using the same approximation as before we obtain

$$\psi_{\alpha_\infty}(\beta) \sim \sum_{i=1}^{d} \ln\left(\frac{1}{\alpha_{\infty,i}}\right)|\beta_i| = \sum_{i=1}^{d} \ln\left(\frac{1}{\alpha_i \exp(-\text{Gain}_\alpha(i))}\right)|\beta_i| = \ln\left(\frac{1}{\alpha}\right)\|\beta\|_1 + \sum_{i=1}^{d} \text{Gain}_\gamma(i)|\beta_i|.$$

Therefore, with a non-uniform Gain_γ, the largest entries will be less likely to be recovered.

3.1 Scale of Gain_γ

First, we observe that if the algorithm takes more time to converge, then the Gain_γ will be larger due to the large number of iterates. Moreover, in Even et al. (2023), it is first proven for a regime where the convergence of the algorithm is ensured that the larger the stepsize, the larger is the Gain_γ. Furthermore, a link is made with the *Edge of Stability* regime, which is the narrow window for the stepsize before the divergence of the algorithm in which the convergence of the algorithm cannot be proven theoretically. In fact, we can observe that in this regime, the algorithm bounces before converging, making the Gain_γ arbitrarily large (see Figure 2 right), resulting in significant differences between (S)GD and GF (see Figure 1).

It is also studied the link of stochasticity with the Gain_γ. In fact, it decreases linearly with the batch size, giving an n difference factor between SGD and GD.

This section studied what distinguishes the GF and the (S)GD. However, to discern the difference in behavior of the latter, we need to understand the shape of the Gain_γ.

3.2 Shape of Gain_γ

In Even et al. (2023), they show that the shape mainly depends on the first few iterates of the optimization algorithm. To do so, they first use the approximation $q(x) \underset{x\sim 0}{\sim} x^2$ to reduce the study of the Gain_γ to studying $\sum_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)^2$. Then, they approximate the latter quantity to $\mathbb{E}[\nabla \mathcal{L}_{\mathcal{B}_k}(\beta_0)^2]$, depending mainly on the initial gradient.

We recall the sparse regression context without noise ($y_i = \langle \beta^\star, x_i \rangle$). In this context, it is proven that the initial gradient $\nabla \mathcal{L}(\beta_0)$ is **heterogeneous** for GD, while the initial stochastic gradients are homogeneous, making the shape of Gainγ **homogeneous** for SGD. In particular, for GD, we have larger values in the support of $\beta^\star$, making it more difficult to recover the true signal. Effectively, first we observe that for the full batch, we can use that

$$\nabla \mathcal{L}(\beta_0)^2 = (\beta^\star_{\text{sparse}})^2 + \epsilon \text{ where } \|\epsilon\|_\infty \ll \|\beta^\star_{\text{sparse}}\|_\infty^2,$$

showing the heterogeneity that makes the support penalty for the GD. Finally, for the batch size 1, we have

$$\mathbb{E}_{i_0}\left[\nabla \mathcal{L}_{i_0}(\beta_0)^2\right] = \Theta(\|\beta^\star_{\text{sparse}}\|_2^2 \mathbf{1}),$$

which proves the homogeneity of the SGD. This can be seen experimentally in left and center of Figure 2.

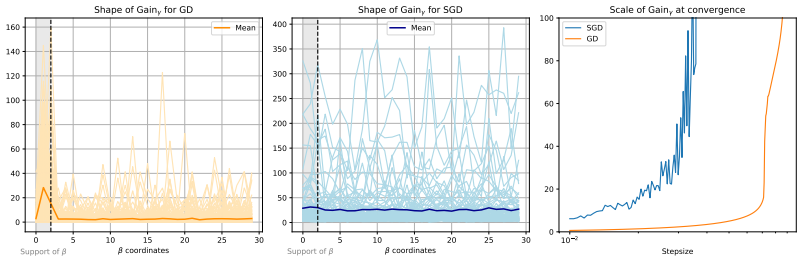

Figure 2: Left: Heterogeneous shape of Gain_γ for the GD. Center: Homogeneous shape of Gain_γ for the SGD. Right: Arbitrarily large Gain_γ on the EoS regime. This proves the dissimilarity between the (S)GD and the GF.

4 Conclusion

Initially, we note that the neural network introduced was a simplified model because the theoretical analysis of more complex networks has not yet been conducted, even with the Gradient Flow simplification. However, these simplified approaches can provide insights into the behavior of more complex networks.

All this theoretically work was presented to fully understand Figure 1. Firstly, we observe that with small stepsizes, the Gain_γ scale is also small, resulting in close solutions between SGD, GD, and GF. However, as the stepsize increases, significant discrepancies emerge between (S)GD, and GF. This difference arises earlier in SGD compared to GD, possibly because stochasticity introduces a linear factor with the stepsize that diminishes the Gain_γ of GD.

Furthermore, to distinguish between GD and SGD, we examine the shape of the Gain_γ. It has been demonstrated that in GD, it is heterogeneous, penalizing non-null components more heavily, thus making it more challenging to recover the true parameters. This explains why the test loss increases. Conversely, in SGD, we achieve better results due to the homogeneity of the Gain_γ, resulting in a minimization problem of a roughly balanced weighted ℓ_1-norm, resembling the uniform ℓ_1-norm.

It is worth noting that the examination of the Gain_γ's shape was only conducted for a batch size of 1 out of n. As a result, there are no findings regarding mini-batch algorithms.

References

Even, M., Pesme, S., Gunasekar, S., and Flammarion, N. (2023). (s)gd over diagonal linear networks: Implicit regularisation, large stepsizes and edge of stability.

Woodworth, B., Gunasekar, S., Savarese, P., Moroshko, E., Golan, I., Lee, J., Soudry, D., and Srebro, N. (2020). Kernel and rich regimes in overparametrized models.

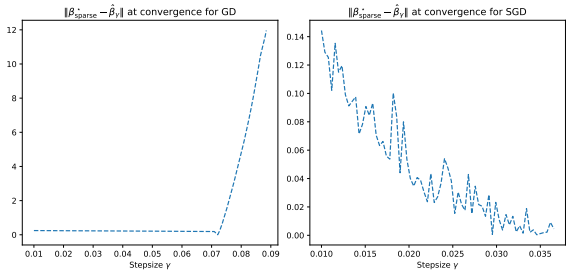

Figure 3: With a macroscopic stepsize, GD fails to recover the true underlying parameter, whereas SGD benefits from it.

A Additional information for Theorem 2.2

The small perturbation term $\widetilde{\beta}_0 \in \mathbb{R}^d$ introduced in Theorem 2.2 is given by $\widetilde{\beta}_0 = \frac{1}{2}(\alpha_+^2 - \alpha_-^2)$ where $q_\pm(x) = \mp 2x - \ln\left((1 \mp x)^2\right)$ and $\alpha_\pm^2 = \alpha^2 \odot \exp\left(-\sum_{k=0}^{\infty} q_\pm\left(\gamma_k \nabla \mathcal{L}_{\mathcal{B}_k}(\beta_k)\right)\right)$.

B (S)GD with macroscopic stepsizes

We will now introduce an alternative perspective to present the same concept as illustrated in Figure 1. Rather than examining the test loss to evaluate the accuracy of each stepsize, we can compare the distance from the true sparse vector $\beta^{\star}_{\text{sparse}}$, as shown in Figure 3.