Implicit Bias of SGD with Structured Noise

Kieran Vaudaux
Section of Mathematics, EPFL
kieran.vaudaux@epfl.ch

Shuailong Zhu
Section of Computer Science, EPFL
shuailong.zhu@epfl.ch

Abstract

The implicit bias of stochastic gradient descent is essential for understanding the training dynamics of neural networks. In this report, we study the continuous version of stochastic gradient descent over the diagonal linear neural network, namely stochastic gradient flow. With the stochastic differential equation modeling, we show the convergence of stochastic gradient flow and the better generalization property of stochastic gradient descent compared to gradient descent, which are the main results of the original paper. The connection between the speed of convergence and the strength of implicit bias is also explained. To better analyze the modeling, we design several experiments to check the consistency of theory and practical training process and to evaluate whether some assumptions of the original paper are plausible.

1 Preliminaries

1.1 Reparameterization of linear regression through diagonal neural network

We focus on a linear regression problem $y = \langle \beta, x \rangle$ with output $(y_1, ..., y_n) \in \mathbb{R}^n$ and inputs $(x_1, ..., x_n) \in (\mathbb{R}^d)^n$, and at least one interpolation parameter β^* exists in the default overparameterized setting ($n < d$). As the linear parameterization doesn't reflect the difference of implicit bias between GD and SGD, we parameterize the regression vector β as β_w non-linearly, where $\beta_w = w_+^2 - w_-^2$ with $w = [w_+, w_-]^T \in \mathbb{R}^{2d}$. This could be thought of as a diagonal linear network of depth 2. We study the quadratic loss,

$$L(w) = L(\beta_w) := \frac{1}{4n} \sum^{n} (\langle \beta_w, x_i \rangle - y_i)^2 = \frac{1}{4n} \sum^{n} (\langle \beta_w - \beta^*, x_i \rangle)^2 \tag{1}$$

1.2 Stochastic Gradient Descent and stochastic gradient flow

Rewrite SGD with noise term. With the quadratic parameterization, the loss term becomes $L(w) = L(\beta_w) := \frac{1}{4n} \sum^{n} (\langle w_+^2 - w_-^2 - \beta^*, x_i \rangle)^2$. The SGD algorithm is:

$$w_{t+1,\pm} = w_{t,\pm} \mp \gamma \langle \beta_w - \beta^*, x_{i_t} \rangle x_{i_t} \odot w_{t,\pm} \quad i_t \sim U(1, n) \tag{2}$$

It is convenient to rewrite as the GD update with a noise term,

$$w_{t+1,\pm} = w_{t,\pm} - \gamma \nabla_{w_\pm} L(w_t) \pm \gamma diag(w_{t,\pm}) X^T \xi_{i_t}(\beta_t) \tag{3}$$

where $\xi_{i_t} = E_{i_t}[\langle \beta_w - \beta^*, x_{i_t} \rangle \boldsymbol{e}_{i_t}] - \langle \beta_w - \beta^*, x_{i_t} \rangle \boldsymbol{e}_{i_t}$, where is a selective canonical basis vector.

The structure of noise term As we are interested in the SDE model of SGD, let us pay attention to the covariance of SGD noise.

$$Cov_{i_t}[\xi_{i_t}(\beta_t)] = \frac{4}{n} diag(L_1(\beta), ..., L_n(\beta)) - \frac{1}{n^2} (\langle \beta_w - \beta^*, x_i \rangle \langle \beta_w - \beta^*, x_j \rangle) \tag{4}$$

where $L_i(\beta) = \frac{1}{4} \langle \beta_w - \beta^*, x_i \rangle)$.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

SDE modeling The author assumes $L_i(\beta) \sim E_{i_t}[L_{i_t}(\beta)]$ and thus $Cov_{i_t}[\xi_{i_t}(\beta_t)] \sim \frac{4}{n}L(\beta) + O(\frac{1}{n^2})$. It leads to the following SDE,

$$dw_{t,\pm} = -\gamma\nabla_{w_\pm}L(w_t)dt \pm 2\sqrt{\gamma n^{-1}L(w_t)}w_{t,\pm} \odot [X^T dB_t] \tag{5}$$

2 Main Results

2.1 Implicit bias of GD

Firstly, we introduce the implicit bias of GD with diagonal neural network. The dynamics of gradient flow is equal to the regime of mirror descent flow with $\phi_\alpha(\beta) = \frac{1}{4}[\sum \beta_i arcsinh(\frac{\beta_i}{2\alpha_i^2}) - \sqrt{\beta_i^2 + 4\alpha^2}]$ as the potential for Bregman divergence.

$$d\nabla\phi_\alpha(\beta_t) = -\nabla L(\beta_t)dt \tag{6}$$

Convergence and implicit bias

- It is showed in appendix of original paper that the loss goes to 0 and the iterates converges towards β_∞.
- Using similar techniques in Lecture 11, we can show: $\beta_\infty = \arg\min_{X\beta=y} D_{\phi_\alpha}(\beta, \beta_0)$, where $\alpha = \sqrt{w_{0,+}^2 w_{0,-}^2}$. If we initialize $w_{0,+} = \alpha$ and $w_{0,-} = \alpha$, then $\beta_0 = 0$. So we have $\beta_\infty^\alpha = argmin_{X\beta=y}\phi_\alpha(\beta)$.
- $\phi_\alpha(\beta)$ is called hyperbolic entropy function, and it is proved in Exercise 11: as $\alpha \to +\infty$, the gradient flow converges to least l2 norm solution while it converges to least l1 norm solution as $\alpha \to 0$.

Some remarks Small initialization of α will leads to a sparse solution, which are known to provide better generalization properties. Large initialization will lead to the kernel regime, which is introduced in Lecture 11.

2.2 Main theorem

The author shows that stochastic gradient descent flow will bias the gradient flow towards solutions which still minimise the hyperbolic entropy with parameter α_∞ which is related to the initialization of weights α and the integral of the SGD loss trajectory.

Theorem 1. For $p \leq 0.5$ and $w_{0,\pm} = \alpha$, we let (w_t) follow the stochastic gradient flow (7) with step size $\gamma \leq 400\left[\ln(\frac{4}{p})\lambda_{\max}\max\left\{||\beta_{l_1}^*||_1 \ln\left(\frac{\|\beta_{l_1}^*\|_1}{\min_i \alpha_i^2}\right)\right\}, \|\alpha\|_2^2\right]^{-1}$. Then with probability $1-p$,

- β_t converges towards a zero training error solution β_∞^α;
- the solution β_∞^α satisfies, $\beta_\infty^\alpha = \arg\min_{s.t. X\beta=y}\phi_{\alpha_\infty}(\beta)$

where $\alpha_\infty = \alpha \odot \exp\left(-2\gamma diag\left(\frac{X^TX}{n}\right)\int_0^\infty L(\beta_s^{SGD})ds\right)$, $\beta_{l_1}^* = \arg\min_{X\beta=y}\|\beta\|_1$.

Sketch of proof : The proof of this theorem requires a large number of intermediate results, which is why we propose here a proof sketch that will give a global understanding of the approach followed in the proof.

- The first step is to show that the iterates from the stochastic gradient flow $(\beta_t)_{t\geq 0}$, coming from the stochastic gradient flows define in Eq.(7), follows a "stochastic continuous mirror descent with time varying potential" defined by: $d\nabla\phi_{\alpha_t}(\beta_t) = -\nabla L(\beta_t)\,dt + \sqrt{\gamma n^{-1}L(\beta_t)}X^T dB_t$ where $\alpha_t = \alpha \odot \exp(-2\gamma\text{diag}(\frac{X^TX}{n})\int_0^t L(\beta_s^{SGD}ds))$, and ϕ_α the hyperbolic entropy.

- Then, for the rest of the proof we need to show that the absolute value of the local martingale $S_t = \sqrt{\gamma}\int_0^t \sqrt{L(\beta_s)\langle \bar{X}^T dB_s, \beta_s - \beta^*_{l_1}\rangle}$ is bounded in probability, with a probability higher then $1-p$ on a event $\mathcal{A} = \left\{\forall t \geq 0, |S_t| \leq a + 2b\gamma\lambda_{\max}\int_0^t L(\beta_s \left(\|\beta_s\|_1^2 + \|\beta^*_{l_1}\|_1^2\right) ds$. This martingale is bounded by a linear function of its quadratic variation for some well chosen constant, a, and slope, b. Once this result is shown the bound of this martingale will be used throughout the proof knowing that it holds with probability at least $1-p$.
- On $\mathcal{A}$, the idea is now to show that the iterates converge to a zero-training error and that the iterates converge to a interpolator β^α_∞. The idea is to follow the same kind of procedure as in the case of the descent mirror in the deterministic framework but this time with a function in the style of the Bregman divergence function but which is this time stochastic. It is at this point that the fact of being in $\mathcal{A}$ will allow us to control this function.
- Finally, the fact that the interpolator β^α_∞ and the gradient $\nabla\phi_{\alpha_\infty}(\beta^\alpha_\infty)$ satisfies the KKT conditions of the minimization problem of the hyperbolic entropy $\phi_\alpha(\beta)$ on the set of the interpolator parameters $\{X\beta = y\}$.

Explanation The most remarkable part of this theorem is that the solution at convergence of stochastic gradient flow minimize the same potential (hyperbolic entropy) for gradient flow with α_∞ strictly smaller than α. As we mentioned in 2.1, smaller α attributes to better generalization.

Interestingly, the scale of α_∞ is controlled by the $\int_0^\infty L(\beta_s)ds$: slower the convergence is, larger the $\int_0^\infty L(\beta_s)ds$ is, the better the generalization property the solution would have, the richer the implicit bias is.

Label Noise As discuss in the paper [1], the slower the iterates converge, the sparser the selected solution will be. Thus, a interesting procedure consist of injecting some artificial label noise, during a period of the training procedure, to slow down the training which may improve the sparsity of the selected interpolator if the iterates converge.

3 Simulation

3.1 Reproduce the results

In Fig 1, it presents the consistency of the SDE model and SGD, which have similar training loss trajectory and convergence point. Also, the test loss for GD at convergence is not as good as SGD, which is consistent with β^α_∞ is not as good as $\beta^{\alpha_\infty}_\infty$ in the aspect of generalization property.

In Fig 2, we first run GD and SGD initialized from α, then we utilize the sum of loss trajectory multiplied by γ to approximate the integral. Then we restart the GD from α_∞ and the test loss at convergence is close to SGD.

Some remarks Every time we choose the γ as large as possible only if SGD converges. Otherwise training process will degrade to kernel regime when the learning rate (step size) decreases which means the SGD, GD and SDE will gradually become the same with the decrease of learning rate.

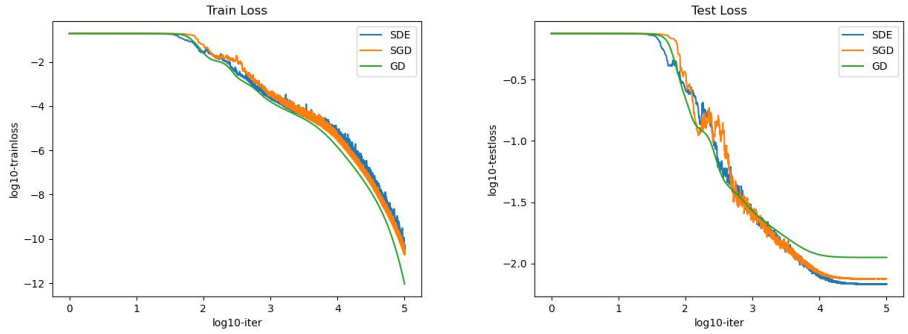
3.2 The simplification of $Cov_{i_t}[\xi_{i_t}(\beta_t)]$ is experimentaly reasonable.

The author assumes $L_i(\beta) \sim E_{i_t}[L_{i_t}(\beta)]$ and simplify the diagonal matrix $diag(L_1(\beta), .., L_n(\beta))$ into $L(\beta)I$. For this part, we want to examine whether this assumption is reasonable through some modified experiments. Thus, we want to compare the SDE accurate form with the SDE in original paper, to see whether they can approximate SGD properly for different distribution of training dataset.

We define **Accurate SDE**:

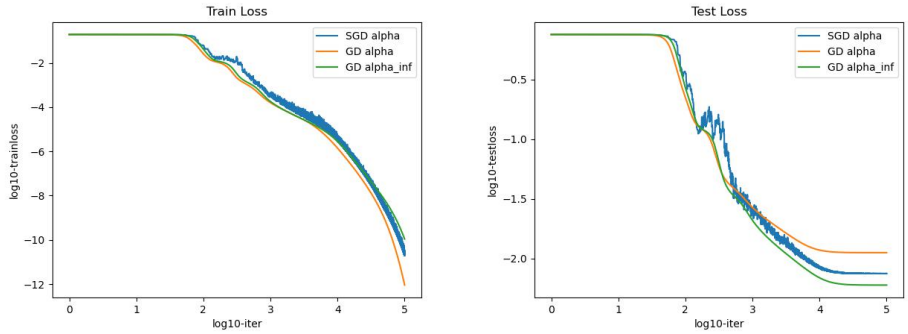
$$dw_{t,\pm} = -\gamma\nabla_{w_\pm}L(w_t)dt \pm 2\sqrt{\gamma n^{-1}}w_{t,\pm} \odot [X^T diag(\sqrt{L_i(\beta)}, ..., \sqrt{L_n(\beta)})dB_t] \quad (7)$$

In Fig. 3, both the SDE and Accurate SDE models the SGD well.



(a) (b)

Figure 1: Validation of SDE model: we only choose one shot SGD and SDE for weights initialized at $\alpha\mathbf{1}$ ($\alpha = 0.01, \gamma = 0.1$). Also, the setting $||\beta||_0 = 5$, $d = 100$ and $n = 40$ will hold in the following.



(a) (b)

Figure 2: Implicit bias from different initializations for SGD and GD: SGD initialized at $\alpha\mathbf{1}$ converges towards the similar point as GD initialized at $\alpha_\infty = \alpha\mathbf{1} \odot exp(-2\gamma diag(\frac{X^TX}{n}) \int_0^\infty L(\beta_s^{SGD} ds))$. ($\alpha = 0.01, \gamma = 0.1$)

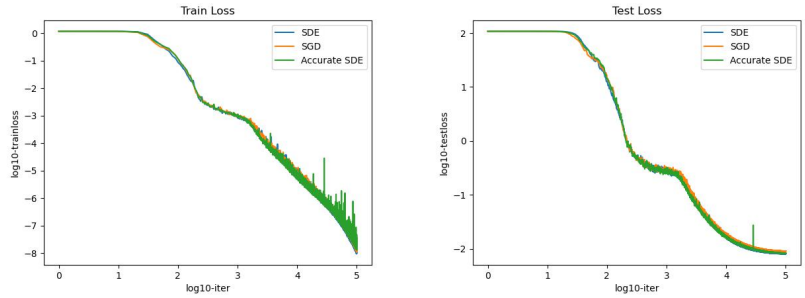
However, the experiments in original paper choose the Gaussian distribution for training dataset, which tend to make each x_i with approximately same norm. Since $L_i(\beta) = \frac{1}{4}\langle \beta_w - \beta^*, x_i \rangle)$, the scale of $L_i(\beta)$ is connected to the norm of x_i. Thus, we sample x_i from Gaussian distributions with different scale, σ_i, to make the norm of x_i different to see whether it will have an influence for the author's simplification.

In Fig. 4, both SDE and Accurate SDE well approximate SGD which seems not influenced by the norm difference of x_i. The reason for the success of the simplification in original paper might be $\sqrt{n^{-1}} w_{t,\pm}$ term with small initialization α make the approximation (replacing $L_i(\beta)$ by $L(\beta)$) not that relevant. Also, if the norm of x_i is too large, we need to adjust the learning rate to be small to converge, which will also eliminate the influence of the norm difference for training dataset.

Some Remarks We also check the consistency when the β is not as sparse as the setting in our report. You could refer to https://github.com/One-punch24/Bias-of-SGF for more results.

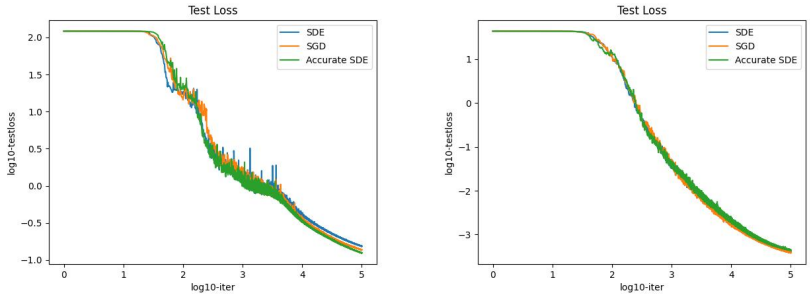
4 Conclusion

In this report, we review the provable results which illustrates why stochastic gradient descent bring better implicit bias than gradient descent. We also try to design numerical experiments to verify its correctness under different settings, which shows the robustness of the stochastic differential equation modeling for stochastic gradient descent.



(a) (b)

Figure 3: Consistency of SDE, Accurate SDE for modeling SGD for training data in uniformly Gaussian distribution: We run SGD, SDE and Accurate SDE initialized at $\alpha\mathbf{1}$ for 10 times and take the average for the loss trajectory ($\alpha = 0.01, \gamma = 0.1$).



(a) (b)

Figure 4: Consistency of SDE, Accurate SDE for modeling SGD for training data in Gaussian distribution with different scale: We sample the x_i from $\sum \frac{1}{n} N(0, start + \frac{end - start}{n})$ and take the average after running SGD, SDE and Accurate SDE for 10 times. For Fig. 4a, $start = 0.01$, $end = 1$, $\gamma = 0.1$, $\alpha = 0.01$; for Fig. 4b, $start = 0.1$, $end = 10$, $\gamma = 0.0025$, $\alpha = 0.01$.

References

[1] Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. *Implicit Bias of SGD for Diagonal Linear Networks: a Provable Benefit of Stochasticity*. 2021. DOI: 10.48550/ARXIV.2106.09524. URL: https://arxiv.org/abs/2106.09524.