Implicit Bias of SGD for Diagonal Linear Networks: a Provable Benefit of Stochasticity

Scott Pesme
EPFL
scott.pesme@epfl.ch

Understanding the implicit bias of training algorithms is of crucial importance in order to explain the success of overparametrised neural networks. We study the dynamics of stochastic gradient descent over diagonal linear networks through its continuous time version, namely stochastic gradient flow. We explicitly characterise the solution chosen by the stochastic flow and prove that it always enjoys better generalisation properties than that of gradient flow. Quite surprisingly, we show that the convergence speed of the training loss controls the magnitude of the biasing effect: the slower the convergence, the better the bias. To fully complete our analysis, we provide convergence guarantees for the dynamics. We also give experimental results which support our theoretical claims. Our findings highlight the fact that structured noise can induce better generalisation and they help explain the greater performances of stochastic gradient descent over gradient descent observed in practice.

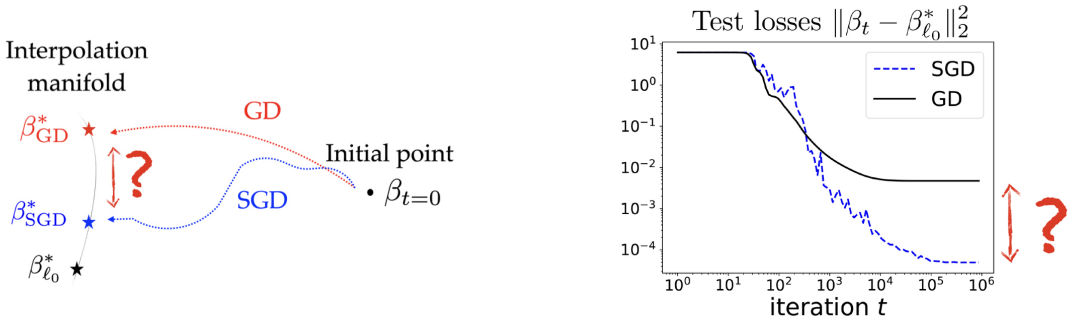

Figure 1: *Left (Drawing)*: For diagonal linear networks, the solutions recovered by SGD and GD differ. *Right*: Sparse regression with $n = 40$, $d = 100$, $\|\beta^*_{\ell_0}\|_0 = 5$, $x_i \sim \mathcal{N}(0, I)$, $y_i = x_i^\top \beta^*_{\ell_0}$. 2-layer diagonal linear network. SGD converges towards a solution which generalises better than GD, the sparsifying effect due to their implicit biases differ by more than an order of magnitude.

Joint work with: Loucas Pillaud-Vivien, Nicolas Flammarion

References

[1] S. Pesme, L. Pillaud-Vivien, N. Flammarion. Implicit Bias of SGD for Diagonal Linear Networks: a Provable Benefit of Stochasticity. *Advances in Neural Information Processing Systems 34 (2021).*