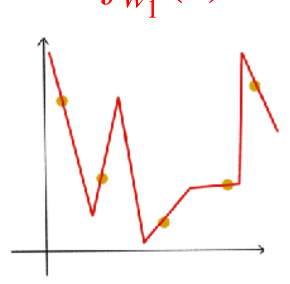
Implicit Bias of SGD for Diagonal Linear Networks:
A Provable Benefit of Stochasticity

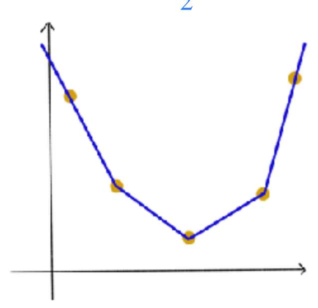
Scott Pesme, Loucas Pillaud-Vivien, Nicolas Flammarion
TML lab

EPFL

Why the concept of implicit bias?



Some ERM ⟵ Training algorithm ⟶ GD SGD etc.



$\mathbb{E}[\ell(f_{w_1^*}(x), y)] \gg 1$ ✗ $\mathbb{E}[\ell(f_{w_2^*}(x), y)] \ll 1$ ✓

How can we accurately model SGD in a continuous way?

SGD

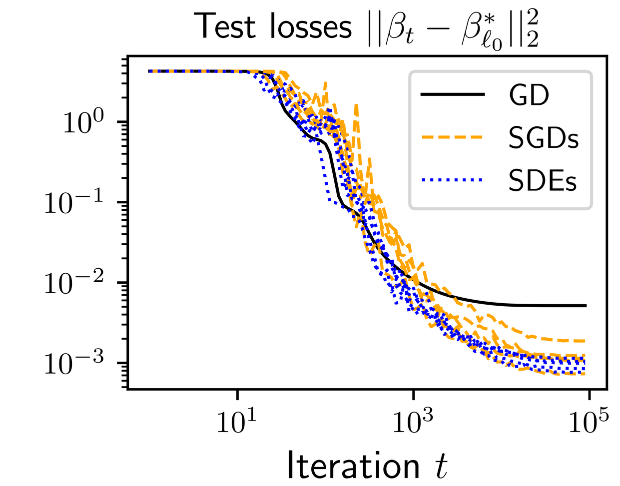
$$u_{t+1} = u_t - \gamma \langle \beta_{w_t} - \beta^*, x_{i_t} \rangle x_{i_t} \odot v_t$$
$$= u_t - \gamma \nabla_u L(w_t) + \underbrace{\gamma v_t \odot [X^\top \xi_{i_t}(w_t)]}_{\text{Zero mean noise}}$$

Stochastic Gradient Flow (SGF)

$$\mathrm{d}u_t = -\nabla_u L(w_t)\mathrm{d}t + \underbrace{2\sqrt{\gamma n^{-1} L(w_t)} v_t \odot [X^\top \mathrm{d}B_t]}_{\text{state dependent !}}$$

(i) matching structure: belongs to $\mathrm{span}(x_1 \odot v, \dots, x_n \odot v)$ (ii) matching covariance $\Sigma_{\mathrm{SGD}}(w)$

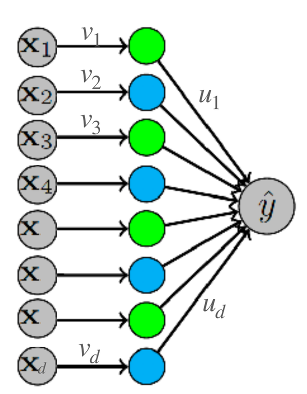
Numerical "validation"



2-layer diagonal linear network

Architecture: Diagonal linear neural network.

Data: $x_1, \dots, x_n \in \mathbb{R}^d$
$y_1, \dots, y_n \in \mathbb{R}$



$f_w(x) = \langle u \odot v, x \rangle$

$w = (u, v) \in \mathbb{R}^{2d}$

Over-parametrised regression task: $d \gg n$

Square-loss: $\min_{w \in \mathbb{R}^{2d}} L(w) = \frac{1}{4n} \sum_{i=1}^{n} (y_i - \langle \underbrace{u \odot v}_{\beta_w}, x_i \rangle)^2$ - Non convex -

$\{\beta_w \in \mathbb{R}^d, L(w) = 0\}$ is a manifold of dim $(d-n)$

Final model is linear but the dynamics is changed

Main result: convergence and implicit bias of the stochastic gradient flow

Assumptions: probability $p \in (0,1)$ and initialisation $u_{t=0} = \alpha \in \mathbb{R}^d, v_{t=0} = 0$. Step-size $\gamma \leq \tilde{O}\left(\frac{1}{\ln(4/p)\lambda_{\max}\|\beta^*_{\ell_1}\|_1}\right)$ where

$$\begin{cases} \lambda_{\max} = \lambda_{\max}(X^\top X/n) \\ \beta^*_{\ell_1} = \underset{\beta \text{ s.t. } X\beta = y}{\mathrm{argmin}} \|\beta\|_1 \end{cases}$$

Result: With probability $1-p$, the Stochastic Gradient Flow (u_t, v_t) is such that:

- *Convergence* → The flow $(\beta_t)_{t\geq 0} = (u_t \odot v_t)_{t \geq 0}$ converges towards a zero-training error solution β^α_∞
- *Implicit Bias* → This solution β^α_∞ satisfies

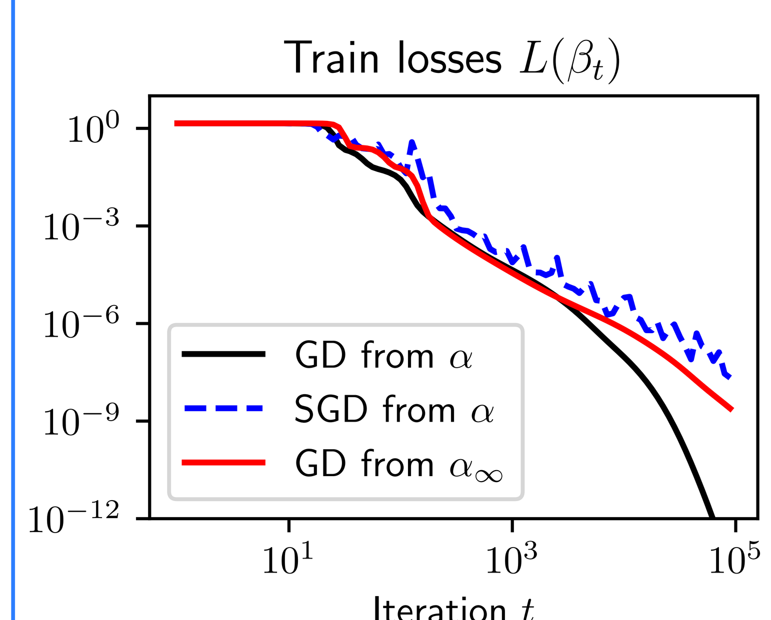
$$\beta^\alpha_\infty = \underset{\beta \in \mathbb{R}^d, \ \langle \beta, x_i \rangle = y_i}{\arg\min} \phi_{\alpha_\infty}(\beta) \quad \text{where} \quad \underbrace{\alpha_\infty}_{\text{"effective" initialisation}} = \alpha \odot \exp\left(-2\gamma \mathrm{diag}\left(\tfrac{X^\top X}{n}\right) \overbrace{\int_0^{+\infty} \underbrace{L(\beta_s)}_{\text{training loss}} \mathrm{d}s}^{\text{stochastic !}}\right) < \overbrace{\alpha}^{\text{initialisation scale}}$$

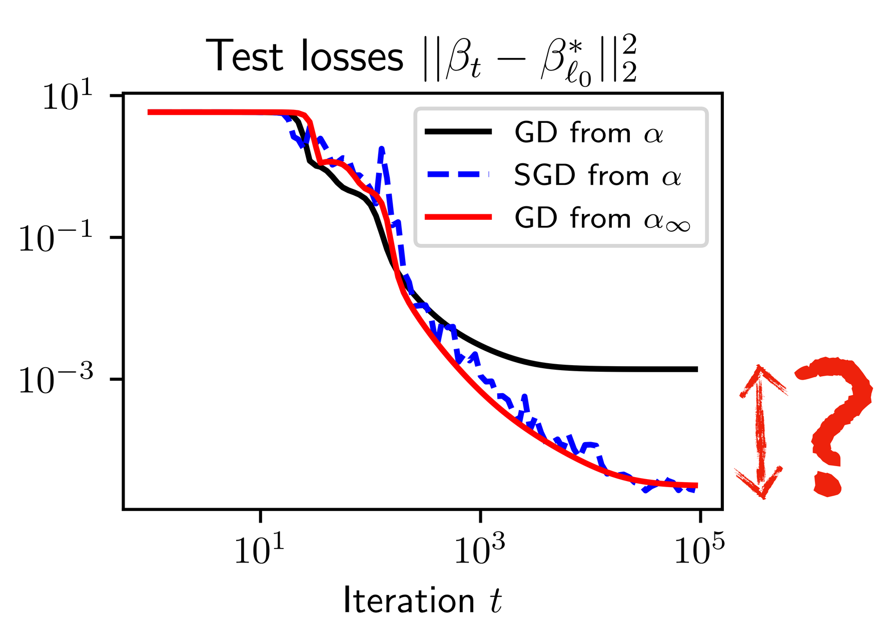
SGD empirically performs better than GD

Gradient flow: $\mathrm{d}w_t = -\nabla_w L(w_t)\,\mathrm{d}t$, $\begin{cases} u_{t=0} = \alpha \in \mathbb{R}^d \\ v_{t=0} = 0 \end{cases} \longrightarrow \beta_{w_{t=0}} = 0$

Implicit bias : $\beta_{w_t} \to \beta^\alpha_\infty = \underset{\beta \in \mathbb{R}^d, \ \langle \beta, x_i \rangle = y_i}{\arg\min} \phi_\alpha(\beta)$ where $\phi_\alpha(\beta) \underset{\alpha \to 0}{\sim} \|\beta\|_1$
(Woodworth et al. 2020) $\qquad \underset{\alpha \to \infty}{\sim} \|\beta\|_2$

What about SGD? Sparse gold model $\beta^*_{\ell_0}$ and $y_i = \langle x_i, \beta^*_{\ell_0} \rangle$





Our paper explains this gap !

Interpretation and observations

GF vs SGF:

- Implicit bias of SGF is the same as GF but with an effective initialisation:

$$\alpha_\infty < \alpha \implies \beta^{\alpha, \mathrm{SGF}}_\infty \text{ is "sparser" than } \beta^{\alpha, \mathrm{GF}}_\infty$$

The slower the convergence, the "better" the bias:

$$\int_0^{+\infty} L(\beta_s)\mathrm{d}s \gg 1 \implies \alpha_\infty \ll \alpha$$

Under additional assumption (boundedness of the iterates):

$$\frac{\alpha_\infty}{\alpha} \underset{\alpha \to 0}{\leq} \left(\frac{\alpha^2}{\|\beta^*_{\ell_1}\|_1}\right)^\zeta \quad \text{for some } \zeta > 0$$

Convergence holds for a fixed step-size:

- This is due to the fact that the noise vanishes at the optimum

Sketch of proof

Stochastic mirror descent with time varying potential:

$$\mathrm{d}\nabla\phi_{\underbrace{\alpha_t}_{\text{stochastic \& time dependent}}}(\beta_t) = \underbrace{-\nabla_\beta L(\beta_t)\ \mathrm{d}t}_{\in\, \mathrm{span}(x_1, \dots, x_n)} + \underbrace{\sqrt{\gamma n^{-1} L(\beta_t)}\ X^\top \mathrm{d}B_t}_{\in\, \mathrm{span}(x_1, \dots, x_n)}$$

$$\alpha_t = \alpha \odot \exp\left(-2\gamma \mathrm{diag}\left(\tfrac{X^\top X}{n}\right) \int_0^t L(\beta_s)\mathrm{d}s\right)$$

Assuming convergence, the KKT conditions immediately give the result:

$$\begin{cases} \nabla\phi_{\alpha_\infty}(\beta^\alpha_\infty) \in \mathrm{span}(x_i) \\ L(\beta^\alpha_\infty) = 0 \end{cases} \xRightarrow{\text{(KKT)}} \beta^\alpha_\infty = \underset{\beta \in \mathbb{R}^d, \ \langle \beta, x_i \rangle = y_i}{\arg\min} \phi_{\alpha_\infty}(\beta)$$

Proving the convergence of the flow $(\beta_t)_{t \geq 0}$ is technically the hardest part:

- Use of appropriate stochastic Lyapunov functions
- Use of martingale concentration inequalities to control the stochastic terms

Woodworth, Blake, et al. "Kernel and rich regimes in overparametrized models.", *COLT* 2020