Deep linear networks for regression are implicitly regularized towards flat minima

NeurIPS 2024

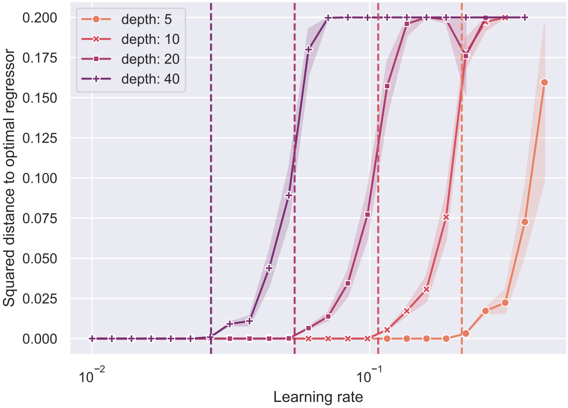
Pierre Marion, Lénaïc Chizat

EPFL

Table of Contents

Table of Contents

GD fails when η exceeds a critical value



Deep linear networks for regression

- Deep linear networks
$$x \mapsto W_L \dots W_1 x \,,$$
with $x \in \mathbb{R}^d$ and parameters $\mathcal{W} = \{\, W_k \in \mathbb{R}^{d_k \times d_{k-1}}\}_{1 \leq k \leq L}$ with $d_L = 1$.
- Regression task: $X \in \mathbb{R}^{n \times d}$, $y \in \mathbb{R}^n$, $\pi^\star$ optimal regressor of minimal norm.
- Mean squared error:
$$R^L(\mathcal{W}) = \frac{1}{n} \| y - X W_1^\top \dots W_L^\top \|_2^2 \,.$$
- **Gradient descent** (GD):
$$\mathcal{W}_{t+1} = \mathcal{W}_t - \eta \nabla R^L(\mathcal{W}_t) \,.$$
- **Notation:** the sharpness $S(\mathcal{W})$ is the largest eigenvalue of the Hessian of R^L.

Where does the critical learning rate value come from?

Damian, Nichani, Lee (2023)

GD implicitly solves

$$\min_{\mathcal{W}} R^L(\mathcal{W}) \quad \text{such that} \quad S(\mathcal{W}) \leq \frac{2}{\eta}\,.$$

- **Interpretation:** GD cannot converge to a minimizer as soon as

$$\eta > \frac{2}{\inf_{\mathcal{W} \in \arg\min(R^L)} S(\mathcal{W})}\,.$$

Where does the critical learning rate value come from?

Damian, Nichani, Lee (2023)

GD implicitly solves

$$\min_{\mathcal{W}} R^L(\mathcal{W}) \quad \text{such that} \quad S(\mathcal{W}) \leq \frac{2}{\eta} \, .$$

- **Interpretation:** GD cannot converge to a minimizer as soon as

$$\eta > \frac{2}{\inf_{\mathcal{W} \in \arg\min(R^L)} S(\mathcal{W})} \, .$$

Theorem

$$\inf_{\mathcal{W} \in \arg\min(R^L)} S(\mathcal{W}) \sim 2La\|\pi^\star\|_2^2 \quad \text{with} \quad a = \left(\frac{\pi^\star}{\|\pi^\star\|}\right)^\top \frac{X^\top X}{n} \frac{\pi^\star}{\|\pi^\star\|} \, .$$

Table of Contents

Our setting

- Mean squared error:
$$R^L(\mathcal{W}) = \frac{1}{n}\|y - XW_1^\top \dots W_L^\top\|_2^2 \,.$$
- Gradient flow (GF):
$$\frac{dW_k}{dt}(t) = -\frac{\partial R^L}{\partial W_k}(t)\,.$$
- Initialization such that $R^L(\mathcal{W}(0)) \leq \frac{1}{n}\|y\|_2^2$ and $\nabla R^L(\mathcal{W}(0)) \neq 0$.

2 questions

- Convergence of gradient flow?
- Structure of the minimizer?

Convergence of GF

Theorem (M. and Chizat, 2024)

The network satisfies the Polyak-Łojasiewicz condition for $t \geq 1$, in the sense that there exists some $\mu > 0$ such that, for $t \geq 1$,

$$\sum_{k=1}^{L} \left\| \frac{\partial R^L}{\partial W_k}(t) \right\|_F^2 \geq \mu(R^L(\mathcal{W}(t)) - R_{\min}) \,.$$

Structure of the minimizer

Corollary

Assume that $32L\sqrt{\varepsilon} \leq 1$ and that the data covariance matrix $\frac{1}{n}X^\top X$ is full rank with smallest (resp. largest) eigenvalue λ (resp. Λ).

Then the gradient flow dynamics converge to a global minimizer $\mathcal{W}^{\mathsf{SI}}$ of the risk, such that

Structure of the minimizer

Corollary

Assume that $32L\sqrt{\varepsilon} \leq 1$ and that the data covariance matrix $\frac{1}{n}X^\top X$ is full rank with smallest (resp. largest) eigenvalue λ (resp. Λ).

Then the gradient flow dynamics converge to a global minimizer $\mathcal{W}^{\mathsf{SI}}$ of the risk, such that

- for $k \in \{1, \dots, L\}$, $\quad \| W_k^{\mathsf{SI}} \|_F^2 - \| W_k^{\mathsf{SI}} \|_2^2 \leq \varepsilon \,,$ (rank-one)
- for $k \in \{1, \dots, L\}$, $\quad \left(\frac{\|\pi^\star\|_2}{2} \right)^{1/L} \leq \sigma_k^{\mathsf{SI}} \leq \left(2\|\pi^\star\|_2 \right)^{1/L} ,$ (low-norm)
- for $k \in \{1, \dots, L-1\}$, $\quad \langle v_{k+1}^{\mathsf{SI}}, u_k^{\mathsf{SI}} \rangle^2 \geq 1 - \dfrac{\varepsilon}{\left(2\|\pi^\star\|_2 \right)^{2/L}} ,$ (alignment)
- $1 \leq \frac{S(\mathcal{W}^{\mathsf{SI}})}{S_{\min}} \leq 4\frac{\Lambda}{\lambda} \,.$ (low-sharpness)

Table of Contents