Three stories on deep linear networks

SIGMA 2024 Workshop, CIRM, Marseille

October 29th, 2024

Pierre Marion, Lénaïc Chizat

Deep linear networks for regression are implicitly regularized towards flat minima

NeurIPS 2024

EPFL

Table of Contents

Table of Contents

Learning rate and sharpness

- Optimization problem

$$\min_{\mathcal{W}\in\mathbb{R}^p} R^L(\mathcal{W})\,.$$

- **Gradient descent** (GD):

$$\mathcal{W}_{t+1} = \mathcal{W}_t - \eta\nabla R^L(\mathcal{W}_t)\,.$$

- Maximal admissible value of η?

Learning rate and sharpness

- Optimization problem

$$\min_{\mathcal{W} \in \mathbb{R}^p} R^L(\mathcal{W}) \, .$$

- **Gradient descent** (GD):

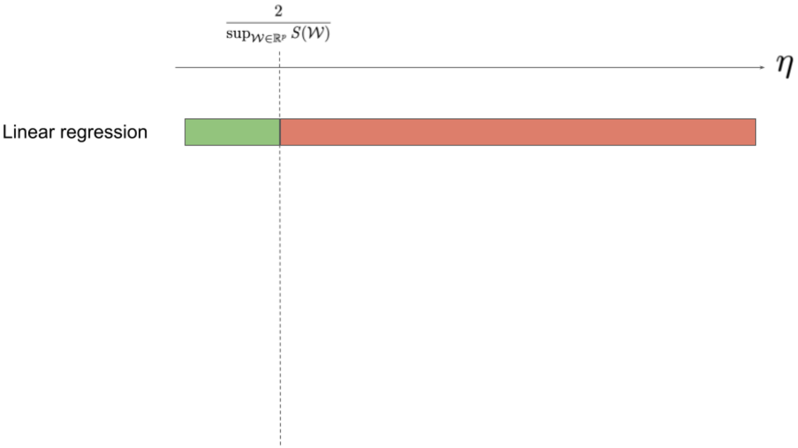
$$\mathcal{W}_{t+1} = \mathcal{W}_t - \eta \nabla R^L(\mathcal{W}_t) \, .$$

- Maximal admissible value of η?
- **Notation:** the sharpness $S(\mathcal{W})$ is the largest eigenvalue of the Hessian of R^L.
- **Convex optimization:** descent lemma for gradient descent (GD) with learning rate η if

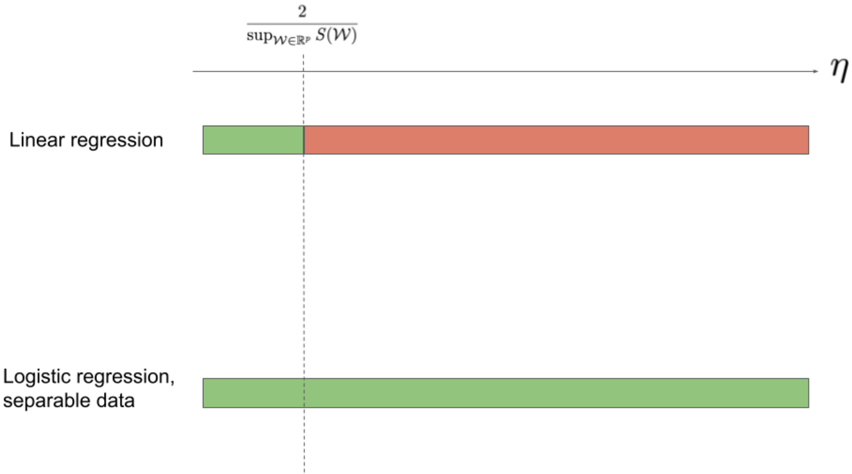
$$\eta < \frac{2}{\sup_{\mathcal{W} \in \mathbb{R}^p} S(\mathcal{W})} \quad \Leftrightarrow \quad \sup_{\mathcal{W} \in \mathbb{R}^p} S(\mathcal{W}) < \frac{2}{\eta} \, .$$

- This is a necessary condition for convergence for a quadratic objective.

Learning rate and sharpness



Learning rate and sharpness



- see Wu, Bartlett, Telgarsky, Yu (2024).

Deep linear networks for regression

- Deep linear networks

$$x \mapsto p^\top W_L \dots W_1 x \,,$$

with $x \in \mathbb{R}^d$, parameters $\mathcal{W} = \{\, W_k \in \mathbb{R}^{d_k \times d_{k-1}} \}_{1 \le k \le L}$, and $p \in \mathbb{R}^{d_L}$ is a fixed vector.

Deep linear networks for regression

- Deep linear networks

$$x \mapsto p^\top W_L \dots W_1 x \,,$$

with $x \in \mathbb{R}^d$, parameters $\mathcal{W} = \{W_k \in \mathbb{R}^{d_k \times d_{k-1}}\}_{1 \le k \le L}$, and $p \in \mathbb{R}^{d_L}$ is a fixed vector.

2 key settings

- Multi-layer perceptron: $d_L = 1$ and $p = 1$.
- Residual network: $d_0 = \cdots = d_L = d$, $W_k \approx I$.

Deep linear networks for regression

- Deep linear networks

$$x \mapsto p^\top W_L \dots W_1 x\,,$$

with $x \in \mathbb{R}^d$, parameters $\mathcal{W} = \{W_k \in \mathbb{R}^{d_k \times d_{k-1}}\}_{1 \leq k \leq L}$, and $p \in \mathbb{R}^{d_L}$ is a fixed vector.

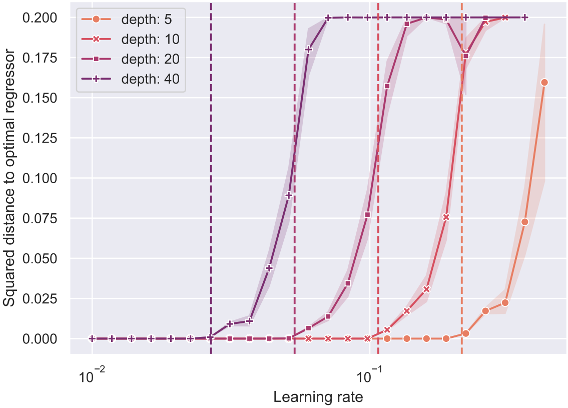
2 key settings

- ▷ Multi-layer perceptron: $d_L = 1$ and $p = 1$.
- ▷ Residual network: $d_0 = \cdots = d_L = d$, $W_k \approx I$.

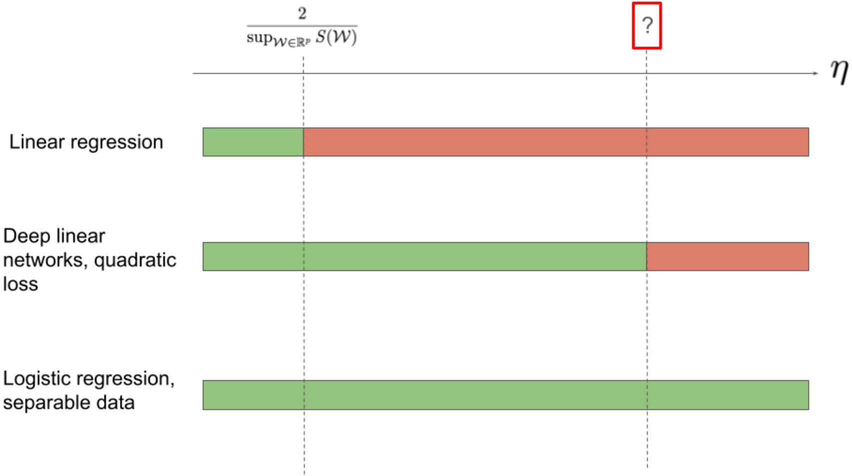
- Regression task: $X \in \mathbb{R}^{n \times d}$, $y \in \mathbb{R}^n$, $\pi^\star$ optimal regressor of minimal norm.
- Mean squared error:

$$R^L(\mathcal{W}) = \frac{1}{n}\|y - XW_1^\top \dots W_L^\top p\|_2^2\,.$$

GD fails when η exceeds a critical value



Learning rate and sharpness



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

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

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

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

Theorem (Mulayoff and Michaeli, 2020; M. and Chizat, 2024)

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

Where does the critical learning rate value come from?

Damian, Nichani, Lee (2023)

GD implicitly solves

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

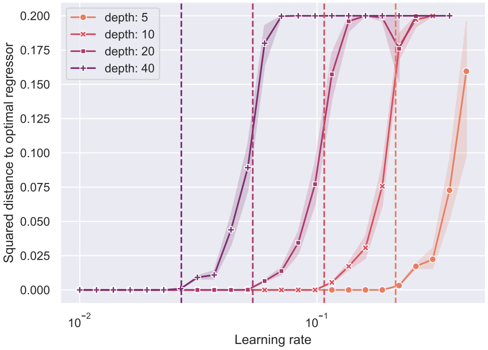
Theorem (Mulayoff and Michaeli, 2020; M. and Chizat, 2024)

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

- GD fails if $\eta > \frac{1}{La\|\pi^\star\|_2^2}$.
- After training to a minimizer, $2La\|\pi^\star\|_2^2 \leq S(\mathcal{W}) \leq \frac{2}{\eta}$.

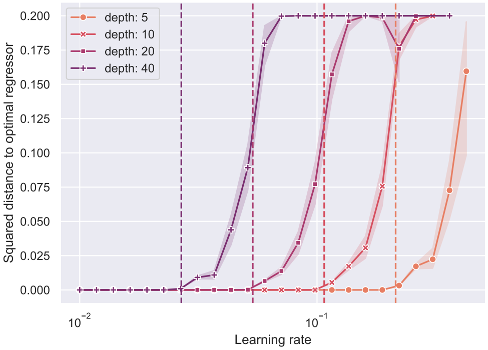
Back to our experiment

- GD fails if $\eta > \frac{1}{La\|\pi^\star\|_2^2}$.

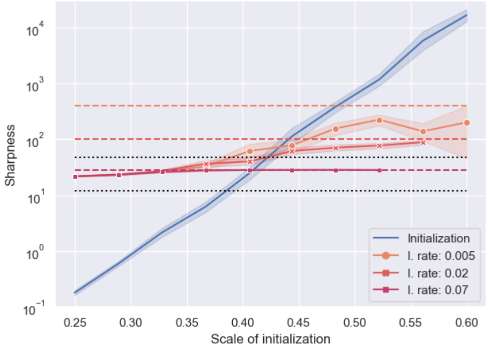


Back to our experiment

- GD fails if $\eta > \frac{1}{La\|\pi^\star\|_2^2}$.

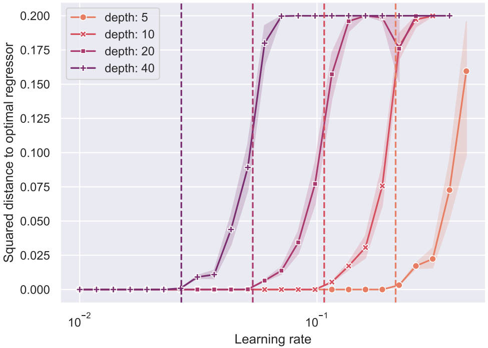


- After training, $2La\|\pi^\star\|_2^2 \leq S(\mathcal{W}) \leq \frac{2}{\eta}$.

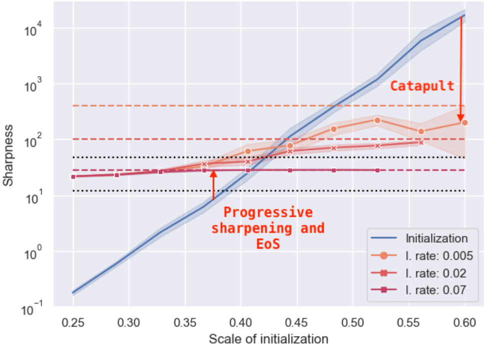


Back to our experiment

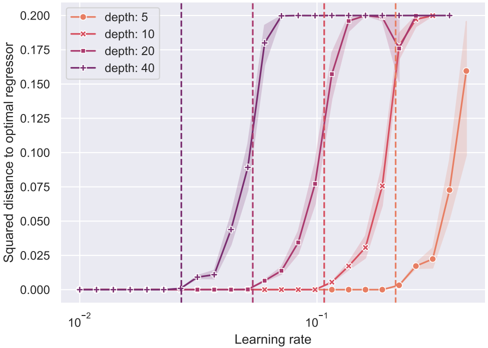
- GD fails if $\eta > \frac{1}{La\|\pi^\star\|_2^2}$.

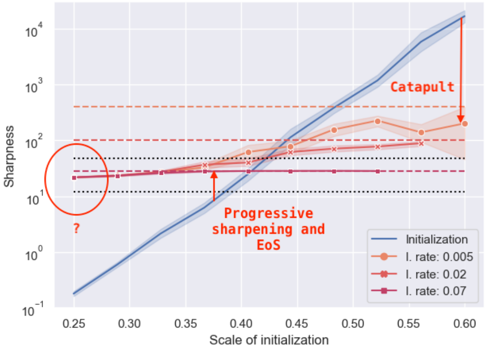


- After training, $2La\|\pi^\star\|_2^2 \leq S(\mathcal{W}) \leq \frac{2}{\eta}$.

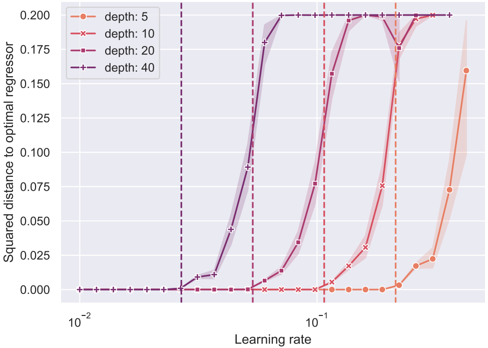


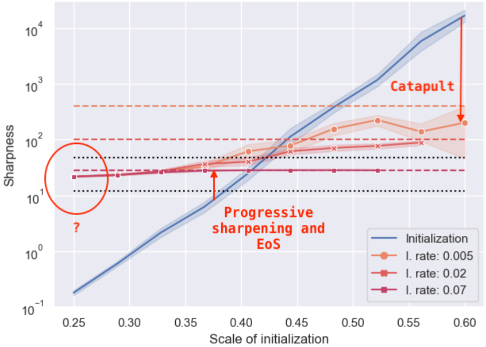
Back to our experiment





Back to our experiment





From a small-scale initialization

- Sharpness does not saturate at $2/\eta$.
- The final sharpness is independent of the learning rate.

Table of Contents

Our setting

- Deep linear networks
$$x \mapsto W_L \dots W_1 x \,,$$
with $x \in \mathbb{R}^d$, parameters $\mathcal{W} = \{ W_k \in \mathbb{R}^{d_k \times d_{k-1}} \}_{1 \le k \le L}$, and $d_L = 1$.
- Mean squared error:
$$R^L(\mathcal{W}) = \frac{1}{n} \| y - X W_1^\top \dots W_L^\top \|_2^2 \,.$$
- Gradient flow (GF):
$$\frac{dW_k}{dt}(t) = -\frac{\partial R^L}{\partial W_k}(t) \,.$$
- Initialization such that $R^L(\mathcal{W}(0)) \le \frac{1}{n}\|y\|_2^2$ and $\nabla R^L(\mathcal{W}(0)) \neq 0$.

2 questions

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

Initialization scale controls the structure of the weights

Define σ_k, u_k, v_k the first singular value, left vector and right vector of W_k, and

$$\varepsilon := 3 \max_{1 \leq k \leq L} \| W_k(0) \|_F^2 + 2 \sum_{k=1}^{L-1} \| W_k(0) W_k^\top(0) - W_{k+1}^\top(0) W_{k+1}(0) \|_2 .$$

Initialization scale controls the structure of the weights

Define σ_k, u_k, v_k the first singular value, left vector and right vector of W_k, and

$$\varepsilon := 3 \max_{1 \leq k \leq L} \| W_k(0) \|_F^2 + 2 \sum_{k=1}^{L-1} \| W_k(0) W_k^\top(0) - W_{k+1}^\top(0) W_{k+1}(0) \|_2 \,.$$

Lemma

The parameters following gradient flow satisfy for any $t \geq 0$ that

- for $k \in \{1, \ldots, L\}$, $\quad \| W_k(t) \|_F^2 - \| W_k(t) \|_2^2 \leq \varepsilon \,,$
- for $j, k \in \{1, \ldots, L\}$, $\quad |\sigma_k^2(t) - \sigma_j^2(t)| \leq \varepsilon \,,$
- for $k \in \{1, \ldots, L-1\}$, $\quad \langle v_{k+1}(t), u_k(t) \rangle^2 \geq 1 - \dfrac{\varepsilon}{\sigma_{k+1}^2(t)} \,.$

Initialization scale controls the structure of the weights

Define σ_k, u_k, v_k the first singular value, left vector and right vector of W_k, and

$$\varepsilon := 3 \max_{1 \leq k \leq L} \| W_k(0) \|_F^2 + 2 \sum_{k=1}^{L-1} \| W_k(0) W_k^\top(0) - W_{k+1}^\top(0) W_{k+1}(0) \|_2 .$$

Lemma

The parameters following gradient flow satisfy for any $t \geq 0$ that

- for $k \in \{1, \dots, L\}$, $\| W_k(t) \|_F^2 - \| W_k(t) \|_2^2 \leq \varepsilon$,
- for $j, k \in \{1, \dots, L\}$, $|\sigma_k^2(t) - \sigma_j^2(t)| \leq \varepsilon$,
- for $k \in \{1, \dots, L-1\}$, $\langle v_{k+1}(t), u_k(t) \rangle^2 \geq 1 - \dfrac{\varepsilon}{\sigma_{k+1}^2(t)}$.

Proof: for any time $t \geq 0$ and any $k \in \{1, \dots, L-1\}$,

$$W_{k+1}^\top(t) W_{k+1}(t) - W_{k+1}^\top(0) W_{k+1}(0) = W_k(t) W_k^\top(t) - W_k(0) W_k^\top(0) .$$

+ computations...

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 $t \geq 1$,

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

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 $t \geq 1$,

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

Beginning of the proof:

$$\frac{\partial R^L}{\partial W_1}(t) = \underbrace{(W_L(t) \dots W_2(t))^\top}_{d_1 \times 1} \underbrace{g^\top}_{1 \times d_0} \,.$$

Therefore

$$\begin{aligned}\left\| \frac{\partial R^L}{\partial W_1}(t) \right\|_F^2 &= \| W_L(t) \dots W_2(t) \|_2^2 \|g\|_2^2 \\ &\geq 4\lambda \| W_L(t) \dots W_2(t) \|_2^2 (R^L(\mathcal{W}(t)) - R_{\min}) \,.\end{aligned}$$

Putting everything together

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 $\mathcal{W}^{\mathrm{SI}}$ of the risk, such that

Putting everything together

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

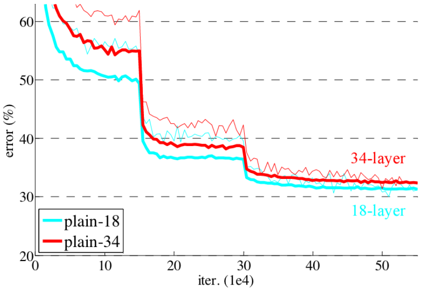
- for $k \in \{1, \dots, L\}$, $\|W_k^{\mathsf{SI}}\|_F^2 - \|W_k^{\mathsf{SI}}\|_2^2 \leq \varepsilon\,,$ (rank-one)
- for $k \in \{1, \dots, L\}$, $\left(\frac{\|\pi^\star\|_2}{2}\right)^{1/L} \leq \sigma_k^{\mathsf{SI}} \leq \left(2\|\pi^\star\|_2\right)^{1/L}\,,$ (low-norm)
- for $k \in \{1, \dots,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

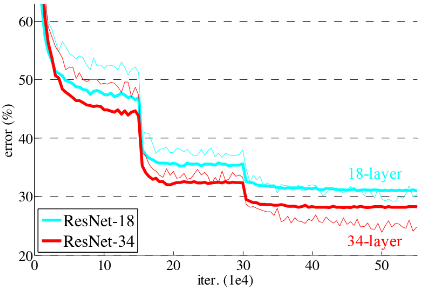
Table of Contents

From multi-layer perceptrons to residual networks

$$h_{k+1} = f(h_k,\, V_{k+1})$$

$$h_{k+1} = h_k + f(h_k,\, V_{k+1})$$





He, Zhang, Ren, Sun (2015)

Linear residual networks

$$h_{k+1} = h_k + V_{k+1} h_k = \underbrace{(I + V_{k+1})}_{=: W_{k+1}} h_k$$

- GF on V_{k+1} initialized at $V(0)$ is equivalent to GF on W_{k+1} initialized at $I + V(0)$.

Linear residual networks

$$h_{k+1} = h_k + V_{k+1} h_k = \underbrace{(I + V_{k+1})}_{=: W_{k+1}} h_k$$

- GF on V_{k+1} initialized at $V(0)$ is equivalent to GF on W_{k+1} initialized at $I + V(0)$.
- Deep linear networks

$$x \mapsto p^\top W_L \dots W_1 x \,,$$

with $x \in \mathbb{R}^d$, parameters $\mathcal{W} = \{ W_k \in \mathbb{R}^{d \times d} \}_{1 \le k \le L}$, and $p \in \mathbb{R}^d$ is fixed.

Linear residual networks

$$h_{k+1} = h_k + V_{k+1} h_k = \underbrace{(I + V_{k+1})}_{=: W_{k+1}} h_k$$

- GF on V_{k+1} initialized at $V(0)$ is equivalent to GF on W_{k+1} initialized at $I + V(0)$.
- Deep linear networks

$$x \mapsto p^\top W_L \dots W_1 x \,,$$

  with $x \in \mathbb{R}^d$, parameters $\mathcal{W} = \{ W_k \in \mathbb{R}^{d \times d} \}_{1 \leq k \leq L}$, and $p \in \mathbb{R}^d$ is fixed.
- Gradient flow (GF):

$$\frac{dW_k}{dt} = -\frac{\partial R^L}{\partial W_k} \,.$$

Linear residual networks

$$h_{k+1} = h_k + V_{k+1} h_k = \underbrace{(I + V_{k+1})}_{=: W_{k+1}} h_k$$

- GF on V_{k+1} initialized at $V(0)$ is equivalent to GF on W_{k+1} initialized at $I + V(0)$.
- Deep linear networks

$$x \mapsto p^\top W_L \dots W_1 x \,,$$

  with $x \in \mathbb{R}^d$, parameters $\mathcal{W} = \{ W_k \in \mathbb{R}^{d \times d} \}_{1 \leq k \leq L}$, and $p \in \mathbb{R}^d$ is fixed.
- Gradient flow (GF):

$$\frac{dW_k}{dt} = -\frac{\partial R^L}{\partial W_k} \,.$$

- Initialization:

$$W_k(0) = I + \frac{s}{\sqrt{Ld}} N_k \,.$$

Zoom on the initialization

$$W_k(0) = I + \frac{s}{\sqrt{Ld}} N_k \,.$$

- N_k: matrices with independent standard Gaussian entries.
- $1/\sqrt{d}$ factor: "right" scaling in the large-width limit.
- $1/\sqrt{L}$ factor: "right" scaling in the large-depth limit.
- s factor: hyperparameter (independent of width and depth).

- On scaling factors, see (for example) Glorot and Bengio (2010); He, Zhang, Ren, Sun (2015); Arpit, Campos, Bengio (2019); Marion, Fermanian, Biau, Vert (2022); Chizat and Netrapalli (2023); Yang, Yu, Zhu, Hayou (2024).

Convergence of GF

Theorem (M. and Chizat, 2024)

There exist $C_1, \dots, C_5 > 0$ depending only on s such that, if $L \geq C_1$ and $d \geq C_2$, then, with probability at least

$$1 - 16 \exp(-C_3 d)\,,$$

if

$$R^L(\mathcal{W}(0)) - R_{\min} \leq \frac{C_4 \lambda^2 \|p\|_2^2}{\Lambda},$$

the gradient flow converges to a global minimizer $\mathcal{W}^{\text{RI}}$ of the risk. Furthermore, the minimizer $\mathcal{W}^{\text{RI}}$ satisfies

$$W_k^{\text{RI}} = I + \frac{s}{\sqrt{Ld}} N_k + \frac{1}{L} \theta_k^{\text{RI}} \quad \text{with} \quad \|\theta_k^{\text{RI}}\|_F \leq C_5\,, \quad 1 \leq k \leq L\,.$$

Concentration of singular values of product of random matrices

Lemma (simplified)

For $u > 0$, with probability at least

$$1 - 8\exp\left(-\frac{du^2}{32s^2}\right),$$

it holds for all θ such that $\max_{1\le k\le L}\|\theta_k\|_2 \le \frac{1}{64}\exp(-2s^2-4u)$ and all $k \in \{1,\dots,L\}$ that

$$\left\|\left(I + \frac{s}{\sqrt{Ld}}N_k + \frac{1}{L}\theta_k\right)\dots\left(I + \frac{s}{\sqrt{Ld}}N_1 + \frac{1}{L}\theta_1\right)\right\|_2 \le 4\exp\left(\frac{s^2}{2}+u\right),$$

and

$$\sigma_{\min}\left(\left(I + \frac{s}{\sqrt{Ld}}N_k + \frac{1}{L}\theta_k\right)\dots\left(I + \frac{s}{\sqrt{Ld}}N_1 + \frac{1}{L}\theta_1\right)\right) \ge \frac{1}{4}\exp\left(-\frac{2s^2}{d}-u\right).$$

Connection with sharpness

Theorem (M. and Chizat, 2024)

The minimizer $\mathcal{W}^{\mathsf{RI}}$ satisfies

$$W_k^{\mathsf{RI}} = I + \frac{s}{\sqrt{Ld}} N_k + \frac{1}{L}\theta_k^{\mathsf{RI}} \quad \text{with} \quad \|\theta_k^{\mathsf{RI}}\|_F \leq C_5 \,, \quad 1 \leq k \leq L \,.$$

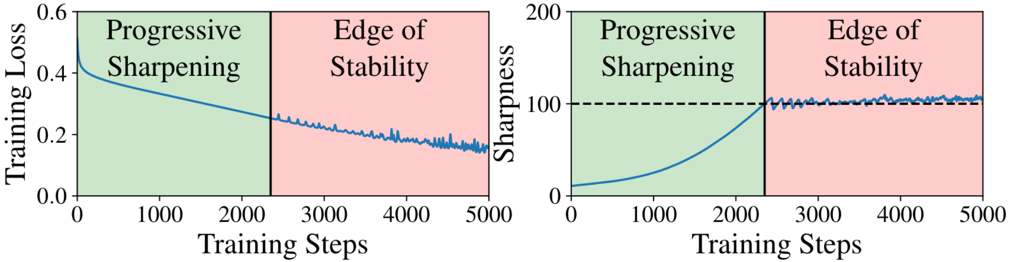
Corollary

If the data covariance matrix $\frac{1}{n}X^\top X$ is full rank, there exists $C > 0$ depending only on s such that the following bounds on the sharpness of the minimizer $\mathcal{W}^{\mathsf{RI}}$ hold:

$$1 \leq \frac{S(\mathcal{W}^{\mathsf{RI}})}{S_{\min}} \leq C\frac{\Lambda}{\lambda} \,.$$

Conclusion: an open problem

Why does the sharpness increase during the early phase of training?



Damian, Nichani, Lee (2023)

Thank you!

Want to know more? arXiv:2405.13456

SCAN ME

pierre.marion@epfl.ch
https://pierremarion23.github.io

All minimizers implement the same optimal regressor

- Mean squared error:
$$R^L(\mathcal{W}) = \frac{1}{n}\|y - XW_1^\top \dots W_L^\top p\|_2^2 \,.$$
- **Assumption:** the covariance matrix $\frac{1}{n}X^\top X \in \mathbb{R}^{d\times d}$ is full-rank, with smallest and largest eigenvalues λ and Λ.
- The linear regression problem of y on X has a unique minimizer $\pi^\star \in \mathbb{R}^d$.
- **Consequence:** all minimizers of $R^L(\mathcal{W})$ are equal in function space to $x \mapsto x^\top \pi^\star$.

Lower bounds on the sharpness of minimizers

Theorem (Mulayoff and Michaeli, 2020; M. and Chizat, 2024)

Let $S_{\min} = \inf_{\mathcal{W} \in \arg\min R^L(\mathcal{W})} S(\mathcal{W})$ and $a := (w^\star/\|w^\star\|)^\top \hat{\Sigma}(w^\star/\|w^\star\|)$. We have

$$S_{\min} \geq 2a\|w^\star\|_2^{2-\frac{1}{L}} \|p\|^{\frac{1}{L}} \sum_{k=1}^{L} \frac{1}{\|W_k\|_F},$$

and

$$2\|w^\star\|_2^{2-\frac{2}{L}} \|p\|^{\frac{2}{L}} La \leq S_{\min} \leq 2\|w^\star\|_2^{2-\frac{2}{L}} \|p\|^{\frac{2}{L}} \sqrt{(2L-1)\Lambda^2 + (L-1)^2 a^2}\,.$$

- The sharpness of minimizers can be arbitrarily high: take any minimizer $\mathcal{W} = (W_1, \dots W_L)$ and consider $\mathcal{W}^C = (CW_1, W_2/C, W_3, \dots, W_L)$. Then

$$S(\mathcal{W}^C) \geq \frac{2\lambda\|\pi^\star\|_2^{2-\frac{1}{L}}}{\|W_2/C\|_F} = \frac{2\lambda\|\pi^\star\|_2^{2-\frac{1}{L}} C}{\|W_2\|_F} \xrightarrow{C\to\infty} \infty.$$

How to lower-bound $\| W_L(t) \dots W_2(t) \|_2$?

$$\left\| \frac{\partial R^L}{\partial W_1}(t) \right\|_F^2 \geq 4\lambda \| W_L(t) \dots W_2(t) \|_2^2 (R^L(\mathcal{W}(t)) - R_{\min}) \,.$$

- Two cases depending on the magnitude of $\sigma_1(t)$.

How to lower-bound $\| W_L(t) \dots W_2(t) \|_2$?

$$\left\| \frac{\partial R^L}{\partial W_1}(t) \right\|_F^2 \geq 4\lambda \| W_L(t) \dots W_2(t) \|_2^2 (R^L(\mathcal{W}(t)) - R_{\min}) \,.$$

- If $\sigma_1(t)$ is "large":

Lemma (reminder)

The parameters following gradient flow satisfy for any $t \geq 0$ that

- for $j, k \in \{1, \dots, L\}$, $\quad |\sigma_k^2(t) - \sigma_j^2(t)| \leq \varepsilon \,.$
- for $k \in \{1, \dots, L-1\}$, $\quad \langle v_{k+1}(t), u_k(t) \rangle^2 \geq 1 - \dfrac{\varepsilon}{\sigma_{k+1}^2(t)} \,.$

How to lower-bound $\| W_L(t) \dots W_2(t)\|_2$?

$$\left\| \frac{\partial R^L}{\partial W_1}(t) \right\|_F^2 \geq 4\lambda \| W_L(t) \dots W_2(t)\|_2^2 (R^L(\mathcal{W}(t)) - R_{\min}) \,.$$

- If $\sigma_1(t)$ is "small":

Assumption (reminder)

- Initialization such that $R^L(\mathcal{W}(0)) \leq \frac{1}{n}\|y\|_2^2$ and $\nabla R^L(\mathcal{W}(0)) \neq 0$.

  - For $t \geq 1$, $\pi(\mathcal{W}(t))$ cannot be too close from 0.
  - Since $\sigma_1(t)$ is small, this implies that $\| W_L(t) \dots W_2(t)\|_2$ is large.

Deep linear networks with small-scale initialization

GF for deep linear networks for regression from a small-scale initialization:

- converges to a global minimum.
- the weights matrices are rank-one and aligned.
- implicit regularization towards small norm and small sharpness.

Some prior work with a similar flavor

- Ji and Telgarsky (2018): aligned and rank-one layers for classification with linearly separable data.
- Saxe et al. (2014, 2019); Lampinen and Ganguli (2019); Gidel et al. (2019); Varre et al. (2023): implicit regularization towards low-rank structure in parameter space for two-layer neural networks.
- Jacot et al. (2021): low-rank saddle-to-saddle dynamics for deep linear networks.

Deep linear networks with residual initialization

GF for deep linear networks for regression from a residual initialization:

- converges when the initial risk is small enough.
- the change to weight matrices is of order $\mathcal{O}(1/L)$.
- the final sharpness can be bounded.

Some prior work with a similar flavor

- Bartlett et al. (2018); Arora et al. (2019); Zou et al. (2020); Sander et al. (2022); Marion et al. (2024): convergence for identity or weight-tied initialization.
- Marion et al. (2022); Zhang et al. (2022): similar concentration bounds for product of random matrices.