Implicit Regularization of Discrete Gradient Dynamics in Deep Linear Neural Networks

Gauthier Gidel[1], **Francis Bach**[2] and **Simon Lacoste-Julien**[1]

[1]Mila, Université de Montréal; [2]SIERRA Project-Team, INRIA and ENS Paris.

Overview

Takeaways

- The choice of the optimization algorithm introduces **biases** that will lead to convergence to specific minimizers of the objective.

  $\boxed{\text{The path matters more than the destination.}}$

- Different optimization algorithms or parametrizations of the model changes the **optimization path**.
- Study the **discrete gradient dynamics** of the training of a **two-layer linear network**. Can be related to matrix factorization [1].
- **Sequentially learns** the solutions of a reduced-rank regression with a gradually increasing rank.

Setting

A Simple Deep Linear Model

Deep linear model: $\hat{\boldsymbol{y}}^d(\boldsymbol{x}) := \boldsymbol{W}_L^\top \cdots \boldsymbol{W}_1^\top \boldsymbol{x}\,,$

Trained with MSE [2]:

$$\min_{\substack{\boldsymbol{W}_l \in \mathbb{R}^{r_{l-1} \times r_l} \\ 1 \le l \le L}} \frac{1}{2n} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W}_1 \cdots \boldsymbol{W}_L\|_2^2 \,.$$

- Thin matrices: **Low rank** constraint, reduced-rank regression,

$$\boldsymbol{W}^{k,*} \in \operatorname*{arg\,min}_{\substack{\boldsymbol{W} \in \mathbb{R}^{p \times d} \\ \text{rank}(\boldsymbol{W}) \le r}} \frac{1}{2n} \sum_{i=1}^{n} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W}\|_2^2 \,.$$

- Large matrices: **overparametrized model**. **Same** expressivity as a linear model but **different** dynamics.

Gradient Dynamics

Discrete gradient dynamics,

$$\boldsymbol{W}_l^{(t+1)} = \boldsymbol{W}_l^{(t)} - \eta \nabla_{\boldsymbol{W}_l} f\big(\boldsymbol{W}_{[L]}^{(t)}\big)$$

Continuous version,

$$\dot{\boldsymbol{W}}_l(t) = -\nabla_{\boldsymbol{W}_l} f\big(\boldsymbol{W}_{[L]}(t)\big)$$

where $\boldsymbol{W}_{[L]}^{(t)} := (\boldsymbol{W}_1^{(t)}, \dots, \boldsymbol{W}_L^{(t)})$.

Assumption

The matrices $\boldsymbol{X}^\top\boldsymbol{X}$ and $\boldsymbol{X}^\top\boldsymbol{Y}$ are close to have **common decomposition**:

$$\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{U}(\boldsymbol{D}_x + \boldsymbol{B})\boldsymbol{U}^\top$$
$$\boldsymbol{X}^\top\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{D}_{xy}\boldsymbol{V}^\top\,,$$

where $\boldsymbol{B}$ is such that $\|\boldsymbol{B}\|_2 \le \epsilon$ and $\boldsymbol{D}_x$, $\boldsymbol{D}_{xy}$ are matrices only with diagonal coefficients.

Sequential Learning of Components

Continuous Case

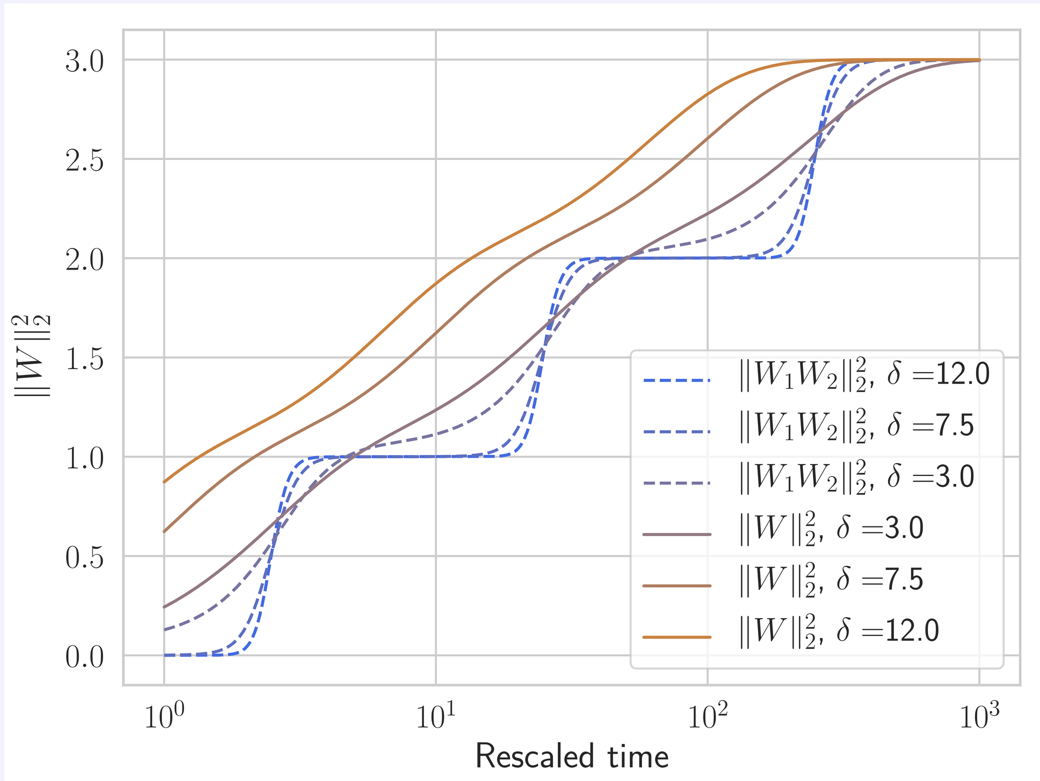
- Can find a **closed form** solution.
- Let $\boldsymbol{W}_1(t)$ and $\boldsymbol{W}_2(t)$ be these solutions.

$$\tfrac{1}{\sigma_k} < t < \tfrac{1}{\sigma_{k+1}} \;\Rightarrow\; \boldsymbol{W}_1(\delta t)\boldsymbol{W}_2(\delta t) \underset{\delta \to \infty}{\to} \boldsymbol{W}^{k,*}$$

  where $\boldsymbol{W}^{k,*}$ is the minimum ℓ_2 norm solution of the reduced-rank-k regression problem.
- Two-layer linear model sequentially find **min-norm low-rank** solutions.
- Notion of *phase transition time*:

$$T_i := \frac{1}{\sigma_i}$$

- Not the case for a one-layer linear model.
- Linear auto-encoder $\boldsymbol{X} = \boldsymbol{Y}$, trace norm is witness of the increasing rank of the solution.



Closed form solutions for a vanishing initialization.

Discrete Case

Why it is interesting:

- Want to understand implicit regularization in gradient based ML.
- In practice *discrete updates*.

Why it is more challenging:

- No *closed form* solution.
- *Infinite* horizon $\to$ cannot consider discrete as an approximation of the continuous.
- New analysis required.

Results:

- Notion of *phase transition time*:

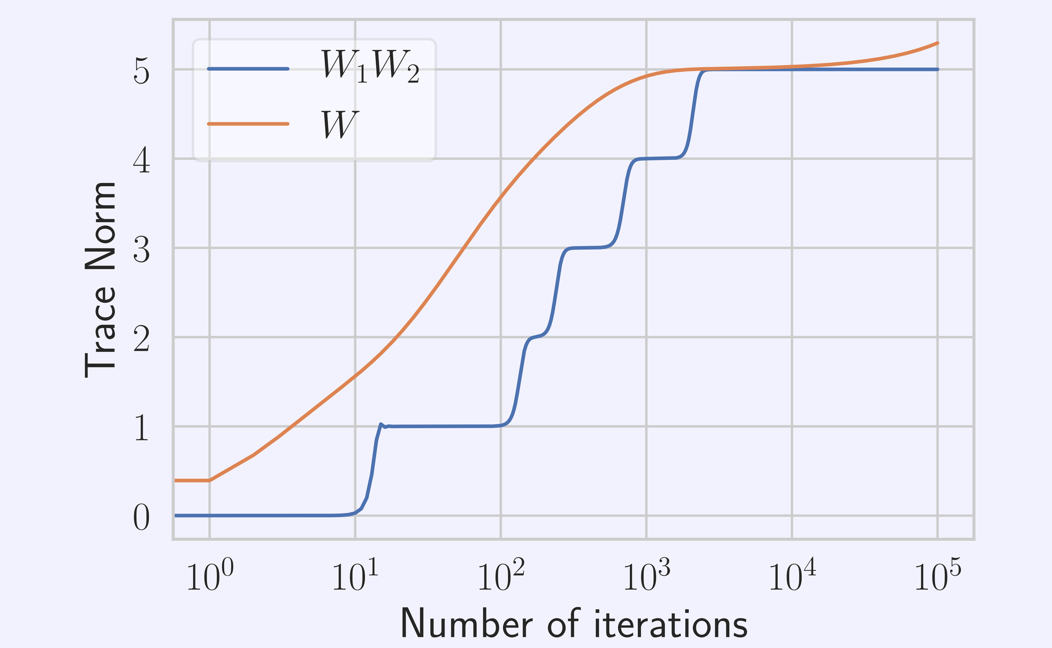
$$T_i := \frac{1}{\eta\sigma_i}$$

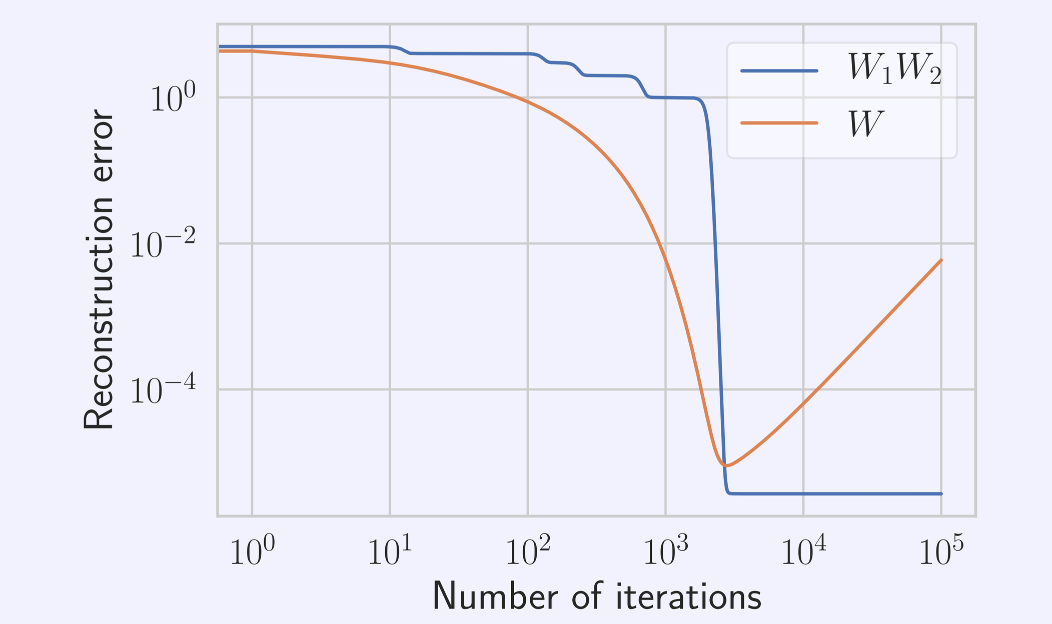
- Similar as the continuous case. (new proof)
- Additional constraints on the step-size.
  - Smaller than a notion of eigen-gap.
  - Smaller than the Lipschitz constant of the gradient.

Experiments and details

Linear auto-encoder

For linear auto encoder $\boldsymbol{X} = \boldsymbol{Y}$ and the trace norm is a witness of the sequentially increasing rank of the solution:





Trace norm and reconstruction errors of $\boldsymbol{W}^{(t)}$ for $L = 1$ and 2 as a function of t.

Detailed Theorems

Continuous Dynamics: if we initialize with,

- $\boldsymbol{W}_1(0) = \boldsymbol{U}\,\mathrm{diag}(e^{-\delta_1}, \dots, e^{-\delta_p})\boldsymbol{Q}$
- $\boldsymbol{W}_2(0) = \boldsymbol{Q}^{-1}\,\mathrm{diag}(e^{-\delta_1}, \dots, e^{-\delta_d})\boldsymbol{V}^\top$

$\boldsymbol{Q}$ is an arbitrary invertible matrix. Then,

$$\boldsymbol{W}_1(t) = \boldsymbol{W}_1^0(t) + \boldsymbol{W}_1^\epsilon(t)$$
$$\boldsymbol{W}_2(t) = \boldsymbol{W}_1^0(t) + \boldsymbol{W}_2^\epsilon(t)$$
$$\boldsymbol{W}_1^0(t) := \boldsymbol{U}\,\mathrm{diag}\big(\sqrt{w_1(t)}, \dots, \sqrt{w_p(t)}\big)\boldsymbol{Q}$$
$$\boldsymbol{W}_2^0(t) := \boldsymbol{Q}^{-1}\,\mathrm{diag}\big(\sqrt{w_1(t)}, \dots, \sqrt{w_d(t)}\big)\boldsymbol{V}^\top$$

where $\|\boldsymbol{W}_i^\epsilon(t)\| \le \epsilon \cdot e^{ct^2}$ and,

$$w_i(t) = \frac{\sigma_i e^{2\sigma_i t - 2\delta_i}}{\lambda_i(e^{2\sigma_i t - 2\delta_i} - e^{-2\delta_i}) + \sigma_i}$$

(σ_i) and (λ_i) are the diagonals of $\boldsymbol{D}_x$ and $\boldsymbol{D}_{xy}$.

Discrete dynamics: under $\epsilon = 0$, we have

$$w_i^{(t)} \ge \frac{w_i^{(0)}}{(\sigma_i - \lambda_i w_i^{(0)})e^{(-2\eta\sigma_i + 4\eta^2\sigma_i^2)t} + w_i^{(0)}\lambda_i}$$

$$w_i^{(t)} \le \frac{w_i^{(0)}}{(\sigma_i - \lambda_i w_i^{(0)})e^{(-2\eta\sigma_i - \eta^2\sigma_i^2)t} + w_i^{(0)}\lambda_i}\,,$$

Differences with the continuous case are in red.

References

[1] S. Gunasekar, B. E. Woodworth, S. Bhojanapalli, B. Neyshabur, and N. Srebro. Implicit regularization in matrix factorization. In *NIPS*, 2017.

[2] A. M. Saxe, J. L. McClelland, and S. Ganguli. A mathematical theory of semantic development in deep neural networks. *arXiv preprint arXiv:1810.10531*, 2018.