Testing a Theory of Wide Networks

Polyak-Łojasiewicz theory predicts loss decrease across epochs

Dario Balboni
dario.balboni@sns.it
Scuola Normale Superiore

Davide Bacciu
davide.bacciu@unipi.it
Università di Pisa

1. The Problem

Much theory is available for very wide networks and networks in the infinite-width limit related to the Neural Tangent Kernel; however it is not clear if such theory is able to explain what happens in real world models.

To answer this question, we measure crucial quantities in the optimization process of realistic models to test three theories related to *convergence*, *conditioning*, and *generalization* of deep networks analyzed under the Polyak-Łojasiewicz (PL) condition.

2. The PL Condition

Definition $h : X \to \mathbb{R}$ is μ-PL iff

$$\frac{1}{2}\|\nabla h(x)\|^2 \geq \mu\,(h(x) - h^*)$$

where $h^* := \inf_{x \in X} h(x)$.

General Strongly convex $\implies$ PL.

Minima Local minima are global:

$$\|\nabla h(x)\| = 0 \implies h(x) = h^*.$$

Uniqueness is not guaranteed.

Convergence is exponential under common first-order methods [KNS16]. For example, when h is L-smooth:

$$h(x_{k+1}) - h^* \leq \left(1 - \frac{\mu}{L}\right)(h(x_k) - h^*),$$

where $x_{k+1} := x_k - \frac{1}{L}\nabla h(x_k)$.

Generalization estimates based on stability are available [CP18]. Generalization improves with a larger μ and with more optimization (larger k), or with a smaller lipschitz constant of h.

3. Relations with Matrix Conditioning

Setting Consider n examples $(x_i, y_i) \in X \times Y$ sampled i.i.d. from an unknown distribution $\mathcal{D}$. A network objective h is a composition of a convex loss function $\mathcal{L} : Y^n \to \mathbb{R}$ with the network evaluation on all points $F : \Theta \to Y^n$: $h = \mathcal{L} \circ F$. Because of PL theory, we are interested in the constants μ and L for h.

Relations with Conditioning It holds that $\mu \propto \lambda_{\min}(K(\theta))$ and $L \propto \lambda_{\max}(K(\theta))$, where $K(\theta) := \nabla F(\theta) \nabla F(\theta)^T$ is the *finite tangent kernel matrix*.

If we train only the last layer we have $\mu \propto \lambda_{\min}(G(\theta))$ and $L \propto \lambda_{\max}(G(\theta))$, where $G_{i,j}(\theta) := \langle F_i(\theta), F_j(\theta) \rangle$ is the kernel matrix of examples passing through the layers.

The conditioning κ of $K(\theta)$, as well as that of $G(\theta)$, decreases exponentially in the depth of infinite-width networks [AAK21], and it is related to PL convergence speed $\gamma := 1 - \mu/L \propto 1 - 1/\kappa$: lower conditioning means faster convergence.

4. Experimental Results

Below measurements relate to overparameterized fully connected networks with ReLU nonlinearities trained under cross-entropy on CIFAR10.

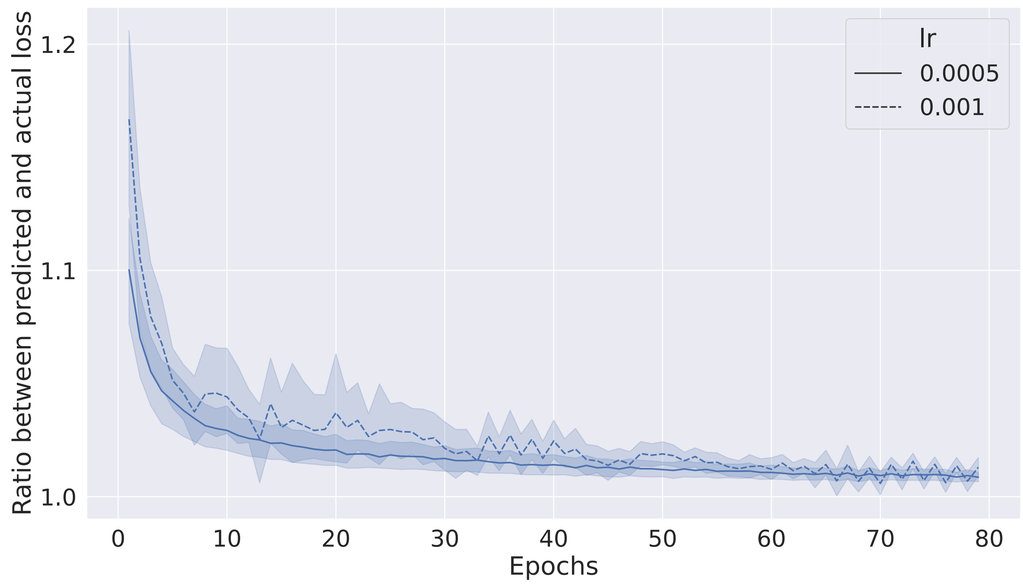

Figure 1: Ratio between predicted and measured loss decrease at single epochs.

Result: The theory is very predictive at later epochs; much less at the start of the optimization.

Future: Better predictions can be used in Neural Architecture Search or in fast model selection.

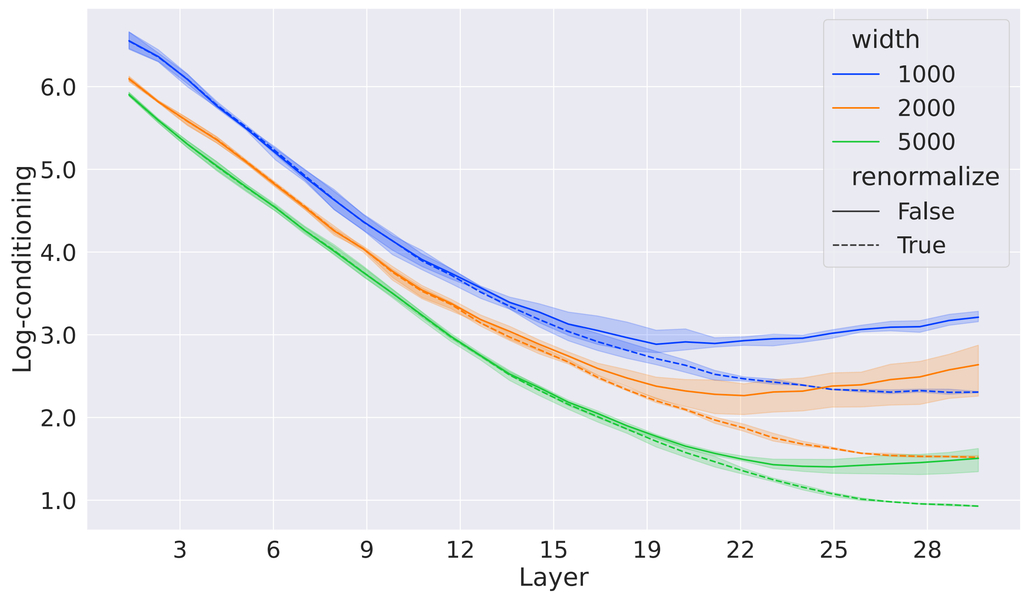

Figure 2: Log-conditioning of $G(\theta)$ for 500 examples at different widths and normalization setup.

Result: Conditioning behaves differently between renormalized and non-renormalized networks, contrary to the theory [AAK21].

Future: Study finite-widths effects on conditioning, as [AAK21] relies on central limit theorem.

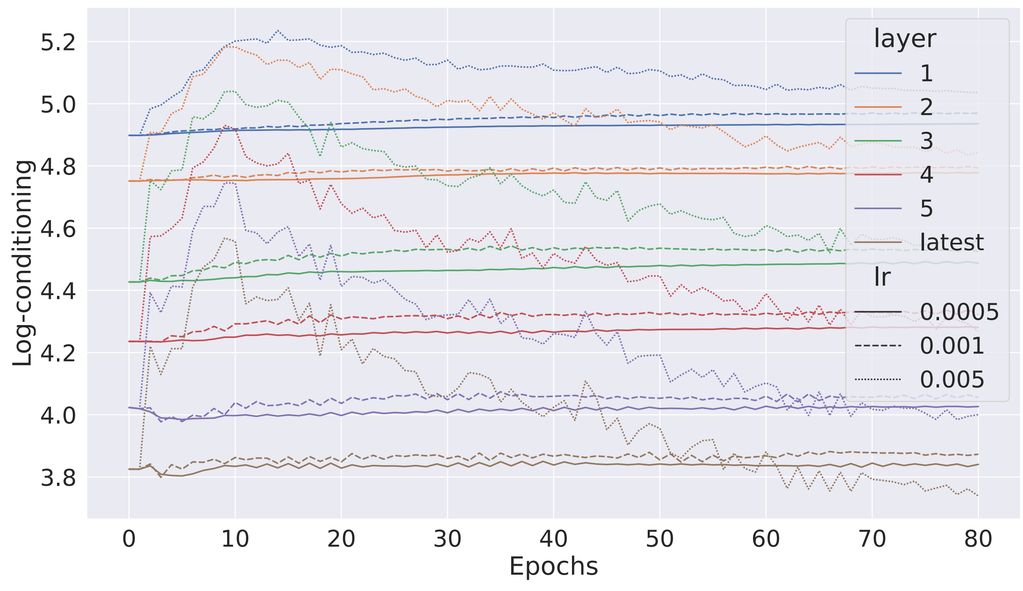

Figure 3: Mean progression of log-conditioning during training.

Result: Conditioning worsens very slowly during training and even improves, contrary to current theories on wide networks.

Future: Understand the underlying phenomena causing a stable conditioning, which simplifies the theoretical analysis of networks.

Main References

[AAK21] N. Agarwal, P. Awasthi, and S. Kale. "A deep conditioning treatment of neural networks". PMLR. 2021.

[CP18] Z. Charles and D. Papailiopoulos. "Stability and generalization of learning algorithms that converge to global optima". 2018.

[KNS16] H. Karimi, J. Nutini, and M. Schmidt. "Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition". 2016.

Want to know more?
Scan the QR Code!