Convergence beyond the over-parameterized regime using Rayleigh quotients

David A. R. Robin
Kevin Scaman
Marc Lelarge

Affiliation

INRIA - École Normale Supérieure de Paris
PSL Research University

Paper link

https://neurips.cc/virtual/2022/poster/54755

Context: Machine Learning, parametric regime

Input set $\mathcal{X}$, output set $\mathcal{Y} \subseteq \mathbb{R}^d$ for some $d \in \mathbb{N}$.
Dataset $\mathcal{D} \subseteq \mathcal{X} \times \mathcal{Y}$, functional loss $\ell : \mathcal{Y}^{\mathcal{X}} \to \mathbb{R}_+$

Least squares $(\mathcal{Y} = \mathbb{R}^k)$: $\ell : f \mapsto \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\|f(x) - y\|_2^2\right]$
Cross entropy $(\mathcal{Y} = \mathbb{R}_+^k)$: $\ell : f \mapsto \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[-\sum_i y_i \log(f_i(x))\right]$

Task: find $f : \mathcal{X} \to \mathcal{Y}$ such that $\ell(f) = 0$.

The Deep Learning tactic:

- Choose $\Theta = \mathbb{R}^m$ a parameter space,
- Parameterize with $F : \Theta \to \mathcal{Y}^{\mathcal{X}}$ to go with $\ell : \mathcal{Y}^{\mathcal{X}} \to \mathbb{R}_+$
- Do gradient flow on $\mathcal{L} : \Theta \to \mathbb{R}_+$, with $\mathcal{L} = \ell \circ F$

Previously in convergence theory: infinite-width NTK

Simplification: finite dataset $\mathcal{X} = [n]$, $\mathcal{Y} = \mathbb{R}$, $\ell(f) = \|f - f^*\|_2^2$.
Parameterization $F : \Theta \to \mathcal{Y}^{\mathcal{X}}$ becomes $F : \mathbb{R}^m \to \mathbb{R}^n$.
Derivative at $\theta \in \mathbb{R}^m$ is a matrix $DF(\theta) \in \mathbb{R}^{n \times m}$.

Neural Tangent Kernel: $K_\theta = DF(\theta)\, DF(\theta)^T \in \mathbb{R}^{n \times n}$.

Prop: If $\exists \mu \in \mathbb{R}_+^*$, $\forall t \in \mathbb{R}_+$, $K_{\theta_t} \succeq \mu$, then $\mathcal{L}(\theta_t) \underset{t \to +\infty}{\longrightarrow} 0$

Proof: By flow def, $-\partial_t \mathcal{L}(\theta) = -\nabla \mathcal{L}_\theta \cdot \partial_t \theta = \nabla \mathcal{L}_\theta \cdot \nabla \mathcal{L}_\theta$
By chain rule on $\mathcal{L} = \ell \circ F$, $\nabla \mathcal{L}_\theta = 2 \cdot DF(\theta)^T (f_\theta - f^*)$, thus

$$-\partial_t \mathcal{L}(\theta) = 4 (f_\theta - f^*)^T K_\theta (f_\theta - f^*) \geq 4\mu \|f_\theta - f^*\|_2^2$$

Therefore $-\partial_t \mathcal{L}(\theta) \geq \kappa \mathcal{L}(\theta)$, thus $\mathcal{L}(\theta_t) \leq \mathcal{L}(\theta_0)\, e^{-\kappa t}$.

That's a Polyak-Łojasiewicz inequality, our proofs are similar.

Kurdyka's desingularizer for Łojasiewicz inequalities

Let $\mathcal{U} \subseteq \Theta$ be a region such that $\mathcal{L} : \Theta \to \mathbb{R}_+$ satisfies the Kurdyka-Łojasiewicz inequality with desingularizer $\varphi : \mathbb{R}_+ \to \mathbb{R}$.

$$\forall \theta \in \mathcal{U}, \quad \mathrm{d}\varphi_{\mathcal{L}(\theta)} \left(\nabla \mathcal{L}_\theta \cdot \nabla \mathcal{L}_\theta \right) \geq \mu$$

If $\theta : \mathbb{R}_+ \to \mathcal{U}$ is a gradient flow of $\mathcal{L}$, then

$$\forall t \in \mathbb{R}_+, \quad \mathcal{L}(\theta_t) \leq \varphi^{-1} \left(\varphi(\mathcal{L}(\theta_0)) - \mu t \right)$$

Ex: $\|\nabla \mathcal{L}\|_2^2 \geq \mathcal{L}$ for $\varphi = \log$, or $\|\nabla \mathcal{L}\|_2^2 \geq \mathcal{L}^2$ for $\varphi(u) = -1/u$

Proof by chain rule.

$$-\partial_t(\varphi \circ \mathcal{L}) = -\mathrm{d}(\varphi \circ \mathcal{L})_\theta \, \partial_t \theta = \mathrm{d}\varphi_{\mathcal{L}(\theta)} \nabla \mathcal{L}_\theta \cdot \nabla \mathcal{L}_\theta \geq \mu$$

Then integrate on the interval $I = [0, t]$. $\square$

$\varphi : \mathbb{R}_+ \to \mathbb{R}$ *pulls back* the affine bound $(I \to \mathbb{R})$ into $(I \to \mathbb{R}_+)$

The problem with definite-NTK assumptions

Recall: with m parameters and n samples, $DF(\theta) \in \mathbb{R}^{n\times m}$

$K_\theta = DF(\theta)\, DF(\theta)^T \in \mathbb{R}^{n\times n}$ has rank $\leq m$

Definite-NTK implies overparameterization
$(K_\theta \succeq \mu > 0) \Rightarrow (m \geq n)$

How do we go to the underparameterized regime?
We can weaken assumption to a Rayleigh quotient bound

Reminder: Rayleigh quotients of bilinear forms

Let $(V, \|\cdot\|_V)$ and $(W, \|\cdot\|_W)$ be two normed vector spaces.
Let $A : V \times W \to \mathbb{R}$ be a bilinear map.

The Rayleigh quotient of A in direction $(x, y) \in V \times W$ is

$$\mathcal{R}\left(A; x, y\right) = \frac{A(x, y)}{\|x\|_V \|y\|_W}$$

If $A : V \times V \to \mathbb{R}$ is a symmetric map, with eigendecomposition $(\lambda_i \in \mathbb{R}_+, v_i \in V)_{i \in [d]}$ orthonormal w.r.t inner product $\langle \cdot, \cdot \rangle$ on V

$$\mathcal{R}\left(A; x, x\right) = \frac{\sum_i \lambda_i \left\langle x, v_i \right\rangle^2}{\sum_i \langle x, v_i \rangle^2}$$

Convex combination of eigenvalues!

Rayleigh bounds are strictly weaker than positive-definiteness.

Kurdyka-Łojasiewicz inequalities by composition

Let $F : \Theta \to \mathcal{F}$ be a differentiable parameterization.
Let $\mathcal{U} \subseteq \Theta$ be a set s.t. $\ell : \mathcal{F} \to \mathbb{R}_+$ satisfies KŁ w. $\varphi : \mathbb{R}_+ \to \mathbb{R}$

$$\forall f \in F(\mathcal{U}), \quad \mathrm{d}\varphi_{\ell(f)} \left(\nabla \ell_f \cdot \nabla \ell_f\right) \geq 1$$

If the Rayleigh quotient of K_θ along $\nabla \ell$ is bounded below on $\mathcal{U}$,

$$\exists \mu \in \mathbb{R}_+^*, \, \forall \theta \in \mathcal{U}, \quad \mathcal{R}\left(K_\theta; \nabla \ell_{F(\theta)}, \nabla \ell_{F(\theta)}\right) \geq \mu$$

Then $\mathcal{L} = (\ell \circ F) : \Theta \to \mathbb{R}_+$ satisfies the KŁ inequality

$$\forall \theta \in \mathcal{U}, \quad \mathrm{d}\varphi_{\mathcal{L}(\theta)} \left(\nabla \mathcal{L}_\theta \cdot \nabla \mathcal{L}_\theta\right) \geq \mu$$

Proof idea: chain rule $\nabla \mathcal{L}_\theta = DF(\theta)^T \nabla \ell_{F(\theta)}$ to make NTK K_θ appear as previously, then use the lower bound assumptions.

Result teaser: Linear-model logistic regression

Input $\mathcal{X} = \mathbb{R}^d$, with $c \in \mathbb{N}^*$ classes. $(\Delta_c = \{p \in \mathbb{R}_+^c \mid \sum_i p_i = 1\})$
Logistic[1] regression with linear models: $F : \mathbb{R}^{c \times d} \to (\mathcal{X} \to \Delta_c)$,

$$F(\theta) : x \mapsto \operatorname{softargmax}(\theta \cdot x)$$

Under multi-class cross-entropy

$$H : f \mapsto \mathbb{E}_x \left[\sum_{i \in [c]} -f^*(x)_i \log \left(f(x)_i \right) \right]$$

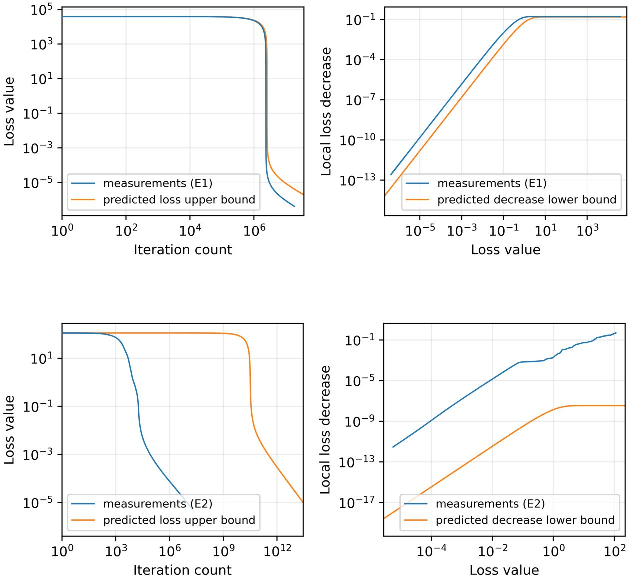
Gradient flows $\theta : \mathbb{R}_+ \to \Theta$ satisfy for all $t \in \mathbb{R}_+$

$$H\left(F(\theta_t)\right) \leq \log \left(\frac{1}{W_0\left(\exp\left(\kappa^2 \varepsilon^2 t - C\right)\right)} \right)$$

With $\varepsilon \in \mathbb{R}_+^*$ a separation margin, $\kappa \in \mathbb{R}_+^*$ an isolation measure, $C \in \mathbb{R}_+^*$ and W_0 is the Lambert function $W_0(x) \exp(W_0(x)) = x$.

[1] $\operatorname{softargmax}(u)_i = e^{u_i} / \sum_j e^{u_j}$

Result teaser: Linear-model logistic regression



Result teaser: Finite-width two-layer neural networks

Input $\mathcal{X} \subseteq \mathbb{R}^d$ compact, $\sigma : \mathbb{R} \to \mathbb{R}$ non-polynomial Lipschitz.
Regression two-layer network: $F : \mathbb{R}^{m \times d} \times \mathbb{R}^m \to (\mathcal{X} \to \mathbb{R})$

$$F(w, a) : x \mapsto \sum_{i \in [m]} a_i \, \sigma(w_i \cdot x)$$

Optimum $f^* : \mathcal{X} \to \mathbb{R}$ is continuous, loss is least-squares

$$\mathcal{L} : \theta \mapsto \mathbb{E}_{x \sim \mathcal{D}} \left[(F(\theta)(x) - f^*(x))^2 \right]$$

Let $\varepsilon \in \mathbb{R}_+^*$ and $\delta \in]0, 1[$
There exists $m \in \mathbb{N}^*$ such that with probability $(1 - \delta)$ over initializations θ_0, all flows $\theta : \mathbb{R}_+ \to \Theta$ with $\theta(0) = \theta_0$ satisfy

$$\mathcal{L}(\theta_t) \underset{t \to +\infty}{\longrightarrow} \eta < \varepsilon$$

Even if $\mathcal{D}$ has infinite support: no over-parameterization here.

Takeaway: Kurdyka-Łojasiewicz + Rayleigh quotients

- Integration of Polyak-Łojasiewicz inequalities works great
  - But they imply linear convergence → implausible for DL?
  - Patch: Replace with Kurdyka-Łojasiewicz inequalities

$$\mathrm{d}\varphi_{\mathcal{L}(\theta)}\left(\nabla\mathcal{L}_\theta \cdot \nabla\mathcal{L}_\theta\right) \geq \mu$$

- Łojasiewicz inequalities (any kind) are very hard to obtain
  - Idea: Proceed by composition (like definite-NTK case)
    (ℓ is KŁ, and F satisfies some property) → ($\ell \circ F$) is KŁ

- Definite-NTK requires overparameterization ($m \geq n$)
  - $K_\theta \in \mathbb{R}^{n\times n}$ has rank $\leq m$ → overparam or rank deficiency
  - Patch: Control *one* Rayleigh quotient, not *all* eigenvalues

- Bonus: Some tools to lower-bound Rayleigh quotients

Convergence beyond the over-parameterized regime using Rayleigh quotients

David A. R. Robin
Kevin Scaman
Marc Lelarge

Affiliation

INRIA - École Normale Supérieure de Paris
PSL Research University

Paper link

https://neurips.cc/virtual/2022/poster/54755