Convergence Rates of Non-Convex Stochastic Gradient Descent Under a Generic Łojasiewicz Condition and Local Smoothness

ICML 2022

Kevin Scaman[1], Cédric Malherbe[2], Ludovic Dos Santos[2]

[1] Inria Paris, DI ENS (work done while at Huawei)
[2] Huawei Noah's Ark lab

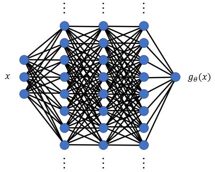
Motivation and setup

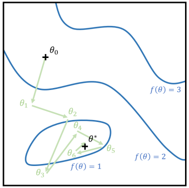
From non-convex SGD to over-paramterized NNs

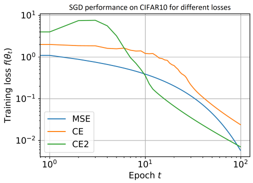
- Training neural networks is usually performed using **non-convex SGD.**
- Recent theoretical analyses show convergence of SGD to a **zero training loss** in the **over-parameterized** setting (i.e. very large number of neurons and layer width).
- In this work, we analyze SGD under **sub-Gaussian gradient noise** to solve

$$\min_{\theta \in \mathbb{R}^d} f(\theta) \triangleq \mathbb{E}\left[\ell(g_\theta(X), Y)\right]$$

where ℓ is a loss function and g_θ is a model parameterized by $\theta \in \mathbb{R}^d$.







Prior works and our contributions

The PL* condition

- Polyak-Łojasiewicz (PL*) condition (Lojasiewicz, 1963; Liu et al., 2020):

$$\forall \theta \in \mathcal{B}(\theta_0, R), \qquad \|\nabla f(\theta)\| \geqslant \sqrt{\mu f(\theta)}\,.$$

- Derived from **uniform conditioning of the NTK** (Jacot et.al., 2018; Liu et al., 2020).
- Limited to **quadratic loss functions** (e.g. MSE).

Our work

- Extends these results to a large class of losses, including **cross entropy**.
- Propose **new conditions** (KL* and SL*) that are more widely applicable.
- Derive **high-probability concentration bounds** for SGD under KL* and SL*.

Convergence of SGD under KL* and SL*

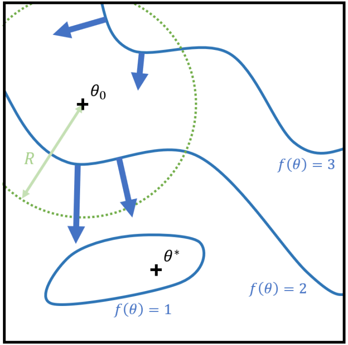
- **Kurdyka-Łojasiewicz (KL*)** condition (Kurdyka, 1998):

$$\forall \theta \in \mathcal{B}(\theta_0, R), \qquad \|\nabla f(\theta)\| \geqslant \varphi\left(f(\theta)\right) .$$

- **Separable-Łojasiewicz (SL*)** condition:

$$\forall \theta \in \mathbb{R}^d, \quad \|\nabla f(\theta)\| \geqslant \phi\left(f(\theta_0) - f(\theta)\right)\psi\left(\|\theta - \theta_0\|\right) .$$

- First term depends on the regularity of the loss.
- Second term depends on the regularity of the model.



Theoretical results

- **High-probability bounds** on the approximation error of SGD.
- Sufficient **control radius** and **convergence time** to reach a given approx. error.

Application to Deep Learning

Assumptions

- Local smoothness of the neural network around initialization;
- Uniform conditioning around initialization, NTK (Liu et al., 2020);
- Lipschitz and smooth loss function w.r.t. its first input.

Properties

Loss function	MSE	HL^2	CE^2	CE	Logistic	Strongly Convex	Convex
Radius	$\Omega(1)$	$\Omega(1)$	$\Omega\left(\ln\left(\frac{1}{\varepsilon}\right)\right)$	$\Omega\left(\ln\left(\frac{1}{\varepsilon}\right)\right)$	$\Omega\left(\ln\left(\frac{1}{\varepsilon}\right)\right)$	$\Omega(1)$	$\Omega(\varepsilon^{-\kappa})$
Time (GD)	$O\left(\ln\left(\frac{1}{\varepsilon}\right)\right)$	$O\left(\ln\left(\frac{1}{\varepsilon}\right)\right)$	$O(\varepsilon^{-1})$	$O(\varepsilon^{-1})$	$O(\varepsilon^{-1})$	$O\left(\ln\left(\frac{1}{\varepsilon}\right)\right)$	$O(\varepsilon^{-1-2\kappa})$
Time (SGD)	$\tilde{O}(\varepsilon^{-2})$	$\tilde{O}(\varepsilon^{-2})$	$\tilde{O}(\varepsilon^{-4})$	$\tilde{O}(\varepsilon^{-4})$	$\tilde{O}(\varepsilon^{-4})$	$\tilde{O}(\varepsilon^{-2})$	$\tilde{O}(\varepsilon^{-4-4\kappa})$

- Convergence of SGD for arbitrary convex losses;
- Flexible approach and robustness of SGD.

Thank you for your attention!