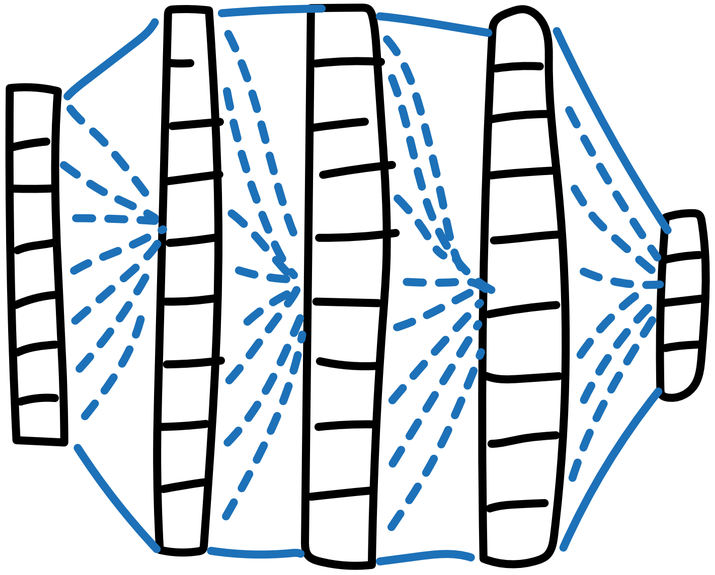
Neural Tangent Kernel

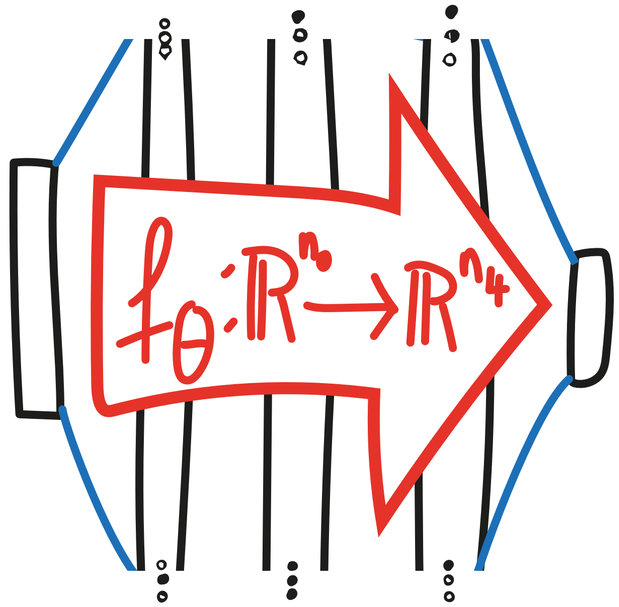
Convergence and Generalization in Neural Networks

Arthur Jacot
arthur.jacot@epfl.ch

Franck Gabriel
franck.gabriel@epfl.ch

Clément Hongler
clement.hongler@epfl.ch

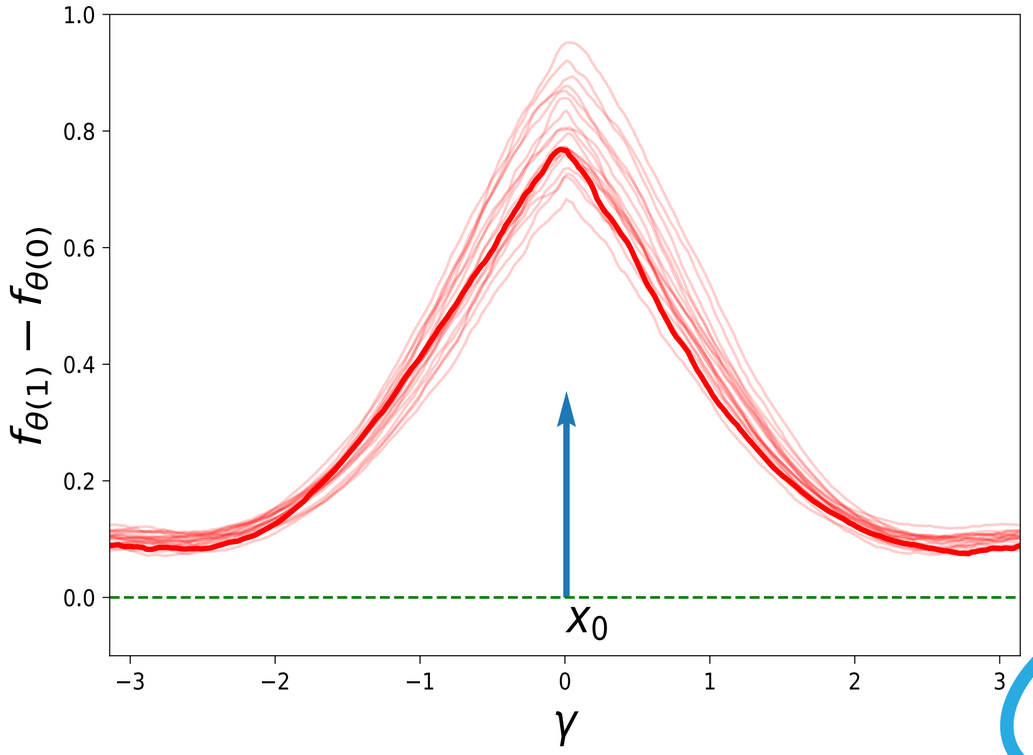




What happens during training?

One step of Gradient Descent

One datapoint x_0



Neural Tangent Kernel:

$$\Theta^{(L)}(x,y) = \sum_{p=1}^{P} \frac{d}{d\theta_p} f_\theta(x) \; \frac{d}{d\theta_p} f_\theta(y)$$

depth

two samples

all parameters

Describes the effect of gradient descent on the network function

In the Infinite width limit:

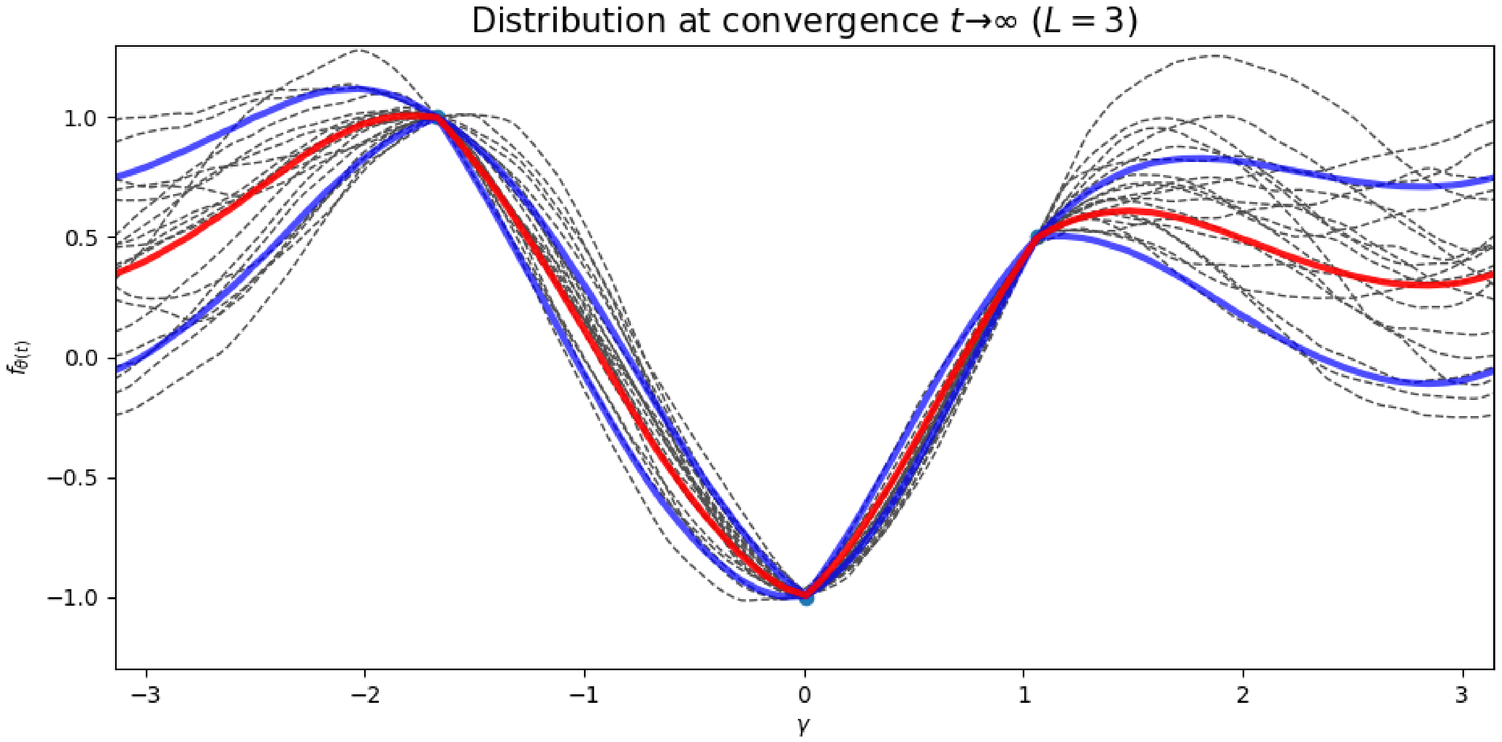
$$\Theta^{(L)}(x,y) \longrightarrow \Theta^{(L)}_{\infty}(x,y)$$

as $n_1, \ldots, n_{L-1} \to \infty$

all hidden Layers

Determines the trajectory of the network function during training

- **Deterministic**
- **Fixed in time**
- **Explicit formula**



Neural Networks	⟺	Kernel methods
Gradient Descent	⟷	Kernel Gradient Descent NTK-regularized gradient
Convergence to a global min.	⟷	Positive definite NTK proved when $\|x_i\|_2 = \|x_j\|_2$
Least-squares loss	⟷	Kernel ridge regression MAP for NTK Gaussian prior

What happens inside a very wide network?

- The activations of the hidden neurons become **independent**
- The parameters and activations **evolve less and less**
- However **all layers learn:**

*The sum of all **microscopic** changes yields a **macroscopic** effect*