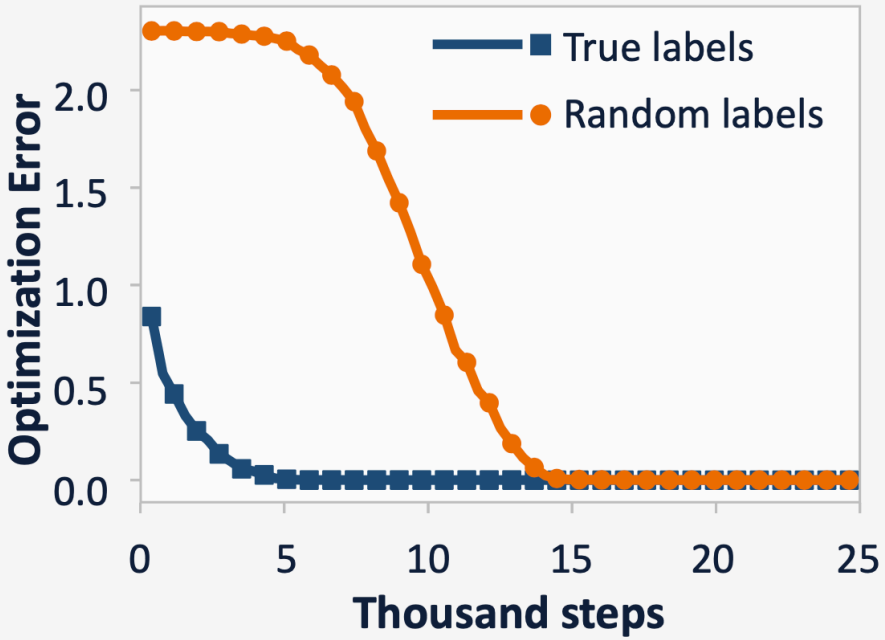
Global convergence of gradient descent

Gradient descent finds global minima

Practice: gradient descent

$$\theta(t+1) \leftarrow \theta(t) - \eta \frac{\partial L(\theta(t))}{\partial \theta(t)}$$



Optimization error **-> 0** for both ***true labels*** and ***random labels*** !

Zhang Bengio Hardt Recht Vinyals 2017

Understanding DL Requires Rethinking Generalization

Global convergence of gradient descent

Theorem (Du et al. ’18, Allen-Zhu et al. ’18, Zou et al ’19) If the width of each layer is poly(n) where n is the number of data. Using random initialization with a particular scaling, gradient descent finds an approximate global minimum in polynomial time.

Gradient Flow: a Kernel Point of View

Gradient Flow: a Kernel Point of View

Gradient Flow: a Kernel Point of View

Gradient Flow: a Kernel Point of View

Gradient Flow: a Kernel Point of View

Gradient Flow: a Kernel Point of View