

Abstract

Context. Recent theory shows that training **wide neural networks** amounts to doing regression with a positive-definite **kernel**.

Contributions. This *lazy training* phenomenon:

- is not intrinsically due to width but to a degenerate relative **scale**
 → depends on **early stopping**, **initialization** and **normalization**
- removes some benefits of depth and may **hinder generalization**

Lazy Training

Setting. Adjust parameters of a differentiable model $h : \mathbb{R}^p \rightarrow \mathcal{F}$ by minimizing a loss $R : \mathcal{F} \rightarrow \mathbb{R}_+$ using gradient flow on the objective

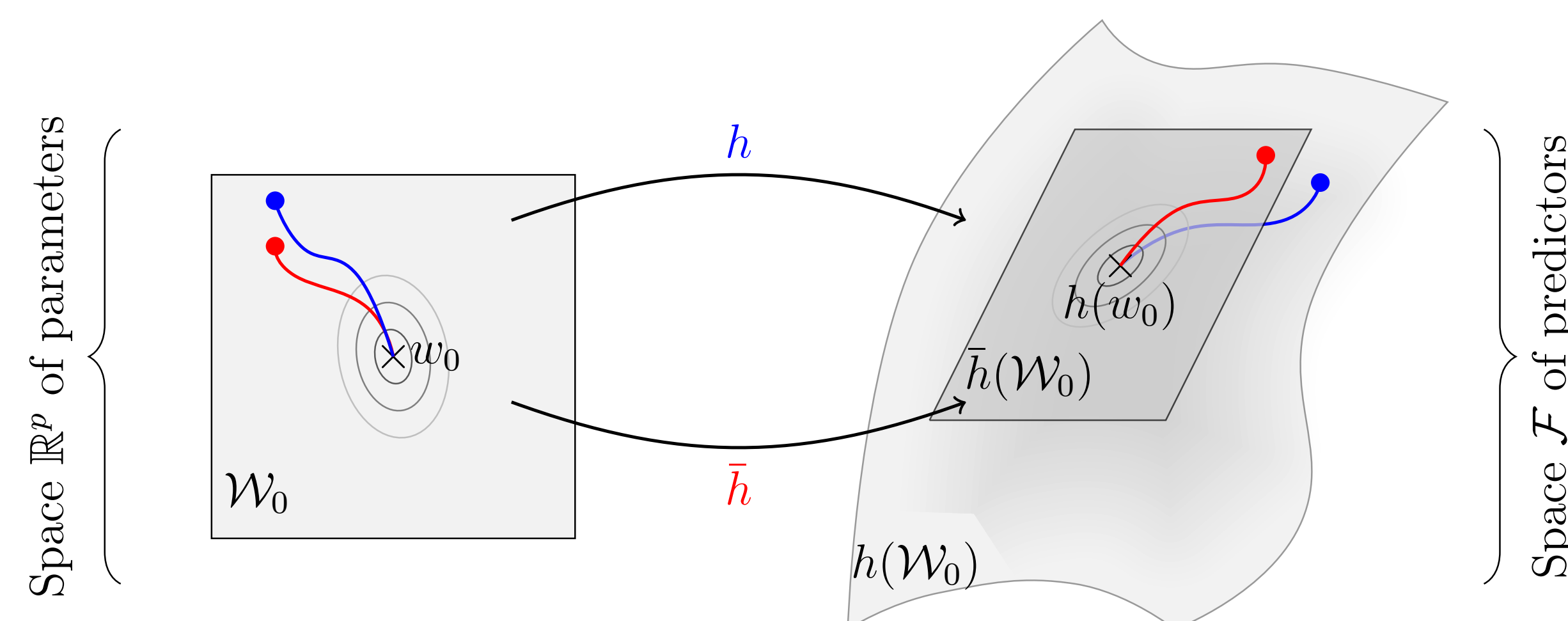
$$F(w) = R(\alpha h(w)) / \alpha^2.$$

- $\mathcal{F}$ is a Hilbert space of predictors, R typically the empirical or population risk, h typically a neural network
- $\alpha > 0$ is a scale, often implicitly present
- gradient flows approximate (stochastic, accelerated) gradient descent

Training paths. For initialization w_0 and stopping time T , let

- $(w_\alpha(t))_{t \in [0, T]}$ be the *original* optimization path
- $(\bar{w}_\alpha(t))_{t \in [0, T]}$ be the *tangent* optimization path, for the tangent model

$$\bar{h}(w) = h(w_0) + Dh(w_0)(w - w_0)$$



Lazy Training (definition)

When the *original* and *tangent* optimization paths are close

Consequences. Lazy training is a type of implicit bias for gradient descent that leads to strong guarantees:

- on optimization speed (theory of convex optimization)
- on generalization (theory of kernel regression)

Lazy Training Theorems

Finite horizon

If $h(w_0) = 0$ and R potentially non-convex then for any $T > 0$,

$$\lim_{\alpha \rightarrow \infty} \sup_{t \in [0, T]} \|\alpha h(w_\alpha(t)) - \alpha \bar{h}(\bar{w}_\alpha(t))\| = 0.$$

Infinite horizon

If $h(w_0) = 0$, and R is strongly convex, then

$$\lim_{\alpha \rightarrow \infty} \sup_{t > 0} \|\alpha h(w_\alpha(t)) - \alpha \bar{h}(\bar{w}_\alpha(t))\| = 0.$$

- over-parameterization is not needed
- see paper for precise statements

When does it occur?

A sufficient criterion. For the square loss $R(y) = \frac{1}{2} \|y - y^*\|^2$ and $\alpha = 1$, the relative difference $\Delta := \|h(w(t)) - \bar{h}(\bar{w}(t))\| / \|y^* - h(w_0)\|$ is controlled by

$$\Delta \lesssim \tilde{t}^2 \cdot \kappa_h(w_0) \quad \text{where} \quad \kappa_h(w_0) := \frac{\|h(w_0) - y^*\| \|D^2 h(w_0)\|}{\|Dh(w_0)\|^2}$$

where $\tilde{t} = t \|Dh(w_0)\|^2$ is the normalized time ($\approx$ iteration number).

Case 1: Rescaled models

For $\alpha > 0$, one has $\kappa_{\alpha h}(w_0) \lesssim \|h(w_0) - y^* / \alpha\|$
 → lazy if $h(w_0)$ small and α large

Case 2: Homogeneous models

If $h(\lambda w) = \lambda^q h(w)$, one has $\kappa_h(\lambda w_0) \lesssim \|h(w_0) - y^* / \lambda^q\|$
 → lazy if $h(w_0)$ small and λ large

Case 3: Wide neural networks

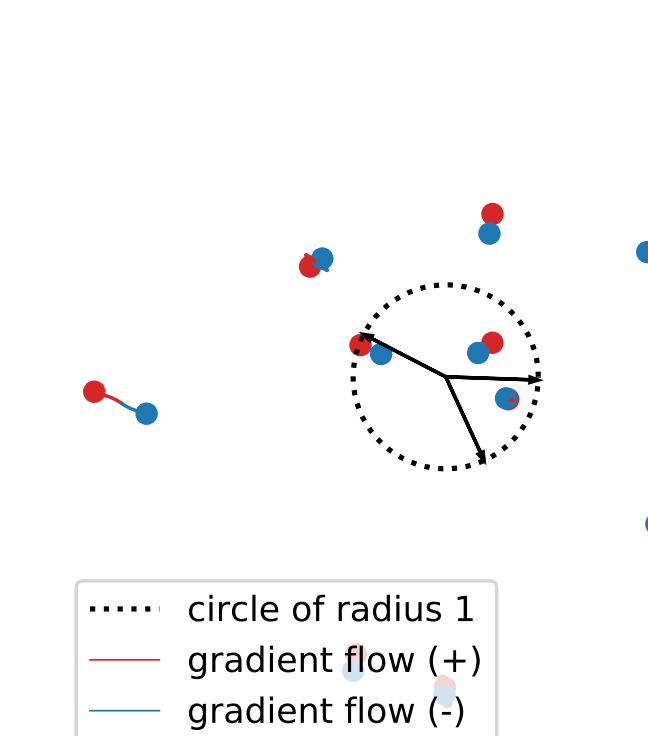
If $h_m(w) = \alpha \sum_{i=1}^m \phi(\theta_i)$ where $w = (\theta_1, \dots, \theta_m)$ are i.i.d. and satisfy $\mathbb{E}\phi(\theta_i) = 0$ (two-layer neural network), then

$$\kappa_{h_m}(w_0) \lesssim m^{-1/2} + (\alpha m)^{-1}$$

- lazy if $\lim_{m \rightarrow \infty} \alpha m = \infty$ (e.g. $\alpha = 1/\sqrt{m}$)
- can be extended to deep networks (Jacot et al.)

Is it desirable in practice?

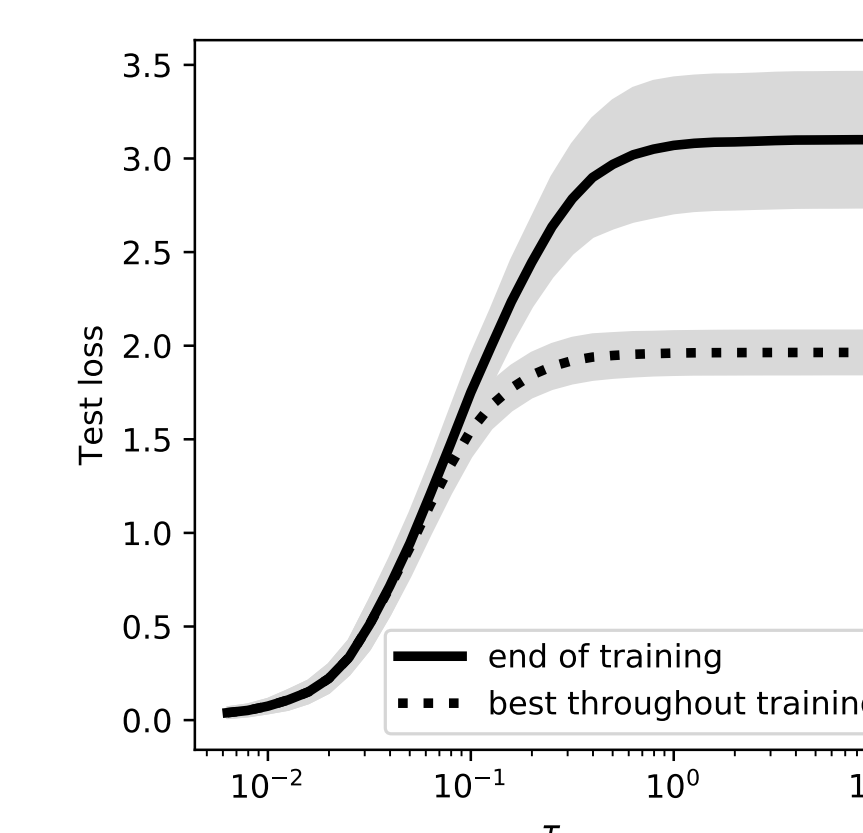
Synthetic experiments. Two-layer ReLU neural network, square loss, initialized with variance τ , best predictor has 3 neurons.



Lazy Training ($\tau = 0.1$)

Non-Lazy Training ($\tau = 2$)

Trajectory of each “hidden” neuron during training (2-D input)

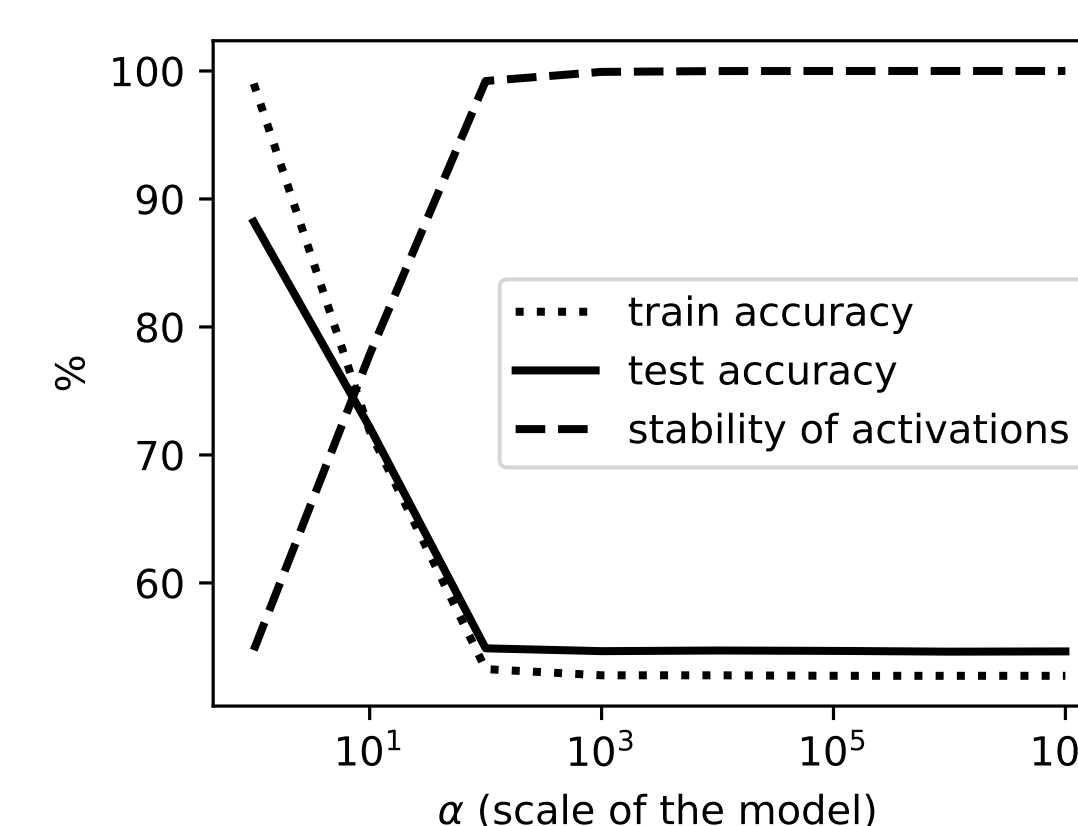


Over-parameterized
(GD on train loss until 0 loss)

Under-parameterized
(SGD on population loss)

Impact of laziness on performance (100-D input)

Image recognition. Does lazy training explain deep learning?



Effect on laziness (VGG11 model) Linear vs. lazy vs. deep models

Model	Train acc.	Test acc.
ResNet wide, linearized	55.0	56.7
VGG-11 wide, linearized	61.0	61.7
Prior features (Oyallon et al.)	-	82.3
Random features (Recht et al.)	-	84.2
VGG-11 wide, standard	99.9	89.7
ResNet wide, standard	99.4	91.0

Theoretical arguments. Neural networks can be superior to kernel/fixed features methods, thanks to their adaptivity (Bach 2017).

Main references

- Jacot et al., *Neural Tangent Kernel: Convergence and Generalization in Neural Networks*. 2018.
- Du et al., *Gradient Descent Provably Optimizes Over-parameterized Neural Networks*. 2018.
- Bach, *Breaking the Curse of Dimensionality with Convex Neural Networks*. 2017.