Two analyses of gradient-based optimization for wide two-layer neural networks

Lénaïc Chizat[*], joint work with Francis Bach[+] and Edouard Oyallon[§]

Nov. 5th 2019 - Theory of Neural Networks Seminar - EPFL

[*]CNRS and Université Paris-Sud [+]INRIA and ENS Paris [§]Centrale Paris

Introduction

Setting

Supervised machine learning

- given input/output training data $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$
- build a function f such that $f(x) \approx y$ for unseen data (x, y)

Gradient-based learning paradigm

- choose a parametric class of functions $f(w, \cdot) : x \mapsto f(w, x)$
- a convex loss ℓ to compare outputs: squared/logistic/hinge...
- starting from some w_0, update parameters using gradients

Example: Stochastic Gradient Descent with step-sizes $(\eta^{(k)})_{k \geq 1}$

$$w^{(k)} = w^{(k-1)} - \eta^{(k)} \nabla_w [\ell(f(w^{(k-1)}, x^{(k)}), y^{(k)})]$$

[Refs]:
Robbins, Monroe (1951). *A Stochastic Approximation Method.*
LeCun, Bottou, Bengio, Haffner (1998). *Gradient-Based Learning Applied to Document Recognition.*

Models

Linear in the parameters

Linear regression, prior/random features, kernel methods:

$$f(w, x) = w \cdot \phi(x)$$

- convex optimization

Models

Linear in the parameters

Linear regression, prior/random features, kernel methods:

$$f(w, x) = w \cdot \phi(x)$$

- convex optimization

Neural networks

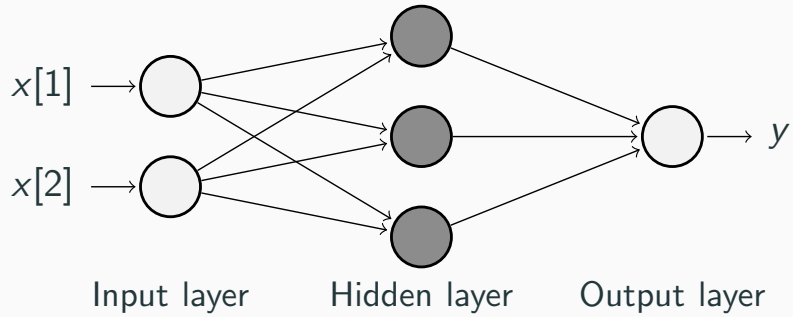
Vanilla NN with activation σ & parameters $(W_1, b_1), \ldots, (W_L, b_L)$:

$$f(w, x) = W_L^T \sigma(W_{L-1}^T \sigma(\ldots \sigma(W_1^T x + b_1) \ldots) + b_{L-1}) + b_L$$

- interacting & interchangeable units/filters
- compositional

Wide two-layer neural networks

Two-layer neural networks



- With activation σ, define $\phi(w_i, x) = c_i \sigma(a_i \cdot x + b_i)$ and

$$f_m(\mathbf{w}, x) = \frac{1}{m} \sum_{i=1}^{m} \phi(w_i, x) \quad \text{with} \quad w_i = (a_i, b_i, c_i) \in \mathbb{R}^p$$

- Estimate the parameters $\mathbf{w} = (w_1, \ldots, w_m)$ by solving

$$\min_{\mathbf{w}} F_m(\mathbf{w}) \quad := \quad \underbrace{R(f_m(\mathbf{w}, \cdot))}_{\text{Empirical/population risk}} \quad + \quad \underbrace{G_m(\mathbf{w})}_{\text{Regularization}}$$

Infinitely wide two-layer networks

- Parameterize the predictor with a probability $\mu \in \mathcal{P}(\mathbb{R}^p)$

$$f(\mu, x) = \int_{\mathbb{R}^p} \phi(w, x) d\mu(w)$$

- Estimate the measure μ by solving

$$\min_{\mu} F(\mu) = \underbrace{R(f(\mu, \cdot))}_{\text{Empirical/population risk}} + \underbrace{G(\mu)}_{\text{Regularization}}$$

- lifted version of "convex" neural networks

[Refs]:

Bengio et al. (2006). *Convex neural networks*.

Adaptivity of neural networks

Goal: Estimate a 1-Lipschitz function $y : \mathbb{R}^d \to \mathbb{R}$ given n iid samples from $\rho \in \mathcal{P}(\mathbb{R}^d)$. Error bound on $\int (\hat{f}(x) - y(x))^2 d\rho(x)$?

- $\tilde{\Omega}(n^{-1/d})$ (curse of dimensionality)

Adaptivity of neural networks

Goal: Estimate a 1-Lipschitz function $y : \mathbb{R}^d \to \mathbb{R}$ given n iid samples from $\rho \in \mathcal{P}(\mathbb{R}^d)$. Error bound on $\int (\hat{f}(x) - y(x))^2 d\rho(x)$?

- $\tilde{\Omega}(n^{-1/d})$ (curse of dimensionality)

Same question, if moreover $y(x) = g(Ax)$ for some $A \in \mathbb{R}^{s \times d}$?

- $\tilde{O}(n^{-1/d})$ for kernel methods (some lower bounds too)
- $\tilde{O}(d^{1/2} n^{-1/(s+3)})$ for 2-layer ReLU networks with weight decay

$\rightsquigarrow$ obtained with $G(\mu) = \int V d\mu$ with $V(w) = \|a\|^2 + |b|^2 + |c|^2$

$\rightsquigarrow$ no *a priori* bound on the number m of units required

$\rightsquigarrow$ connecting theory and practice:

Is it related to the predictor learnt by gradient descent?

[Refs]:
Barron (1993). *Approximation and estimation bounds for artificial neural networks.*
Bach. (2014). *Breaking the curse of dimensionality with convex neural networks.*

Mean-field dynamic and global convergence

Continuous time dynamics

Gradient flow

Initialize $\mathbf{w}(0) = (w_1(0), \ldots, w_m(0))$.

Small step-size limit of (stochastic) gradient descent:

$$\mathbf{w}(t+\eta) = \mathbf{w}(t) - \eta \nabla F_m(\mathbf{w}(t)) \quad \underset{\eta \to 0}{\Rightarrow} \quad \frac{d}{dt}\mathbf{w}(t) = -m \nabla F_m(\mathbf{w}(t))$$

Measure representation

Corresponding dynamics in the space of probabilities $\mathcal{P}(\mathbb{R}^p)$:

$$\mu_{t,m} = \frac{1}{m} \sum_{i=1}^{m} \delta_{w_i(t)}$$

Technical note: in what follows $\mathcal{P}_2(\mathbb{R}^p)$ is the *Wasserstein space*

Many-particle / mean-field limit

Theorem

Assume that $w_1(0), w_2(0), \ldots$ *are such that* $\mu_{0,m} \to \mu_0$ *in* $\mathcal{P}_2(\mathbb{R}^p)$ *and some regularity. Then* $\mu_{t,m} \to \mu_t$ *in* $\mathcal{P}_2(\mathbb{R}^p)$*, uniformly on* $[0, T]$*, where* μ_t *is the unique Wasserstein gradient flow of* F *starting from* μ_0*.*

Wasserstein gradient flows are characterized by

$$\partial_t \mu_t = -\mathrm{div}(-\nabla F'_{\mu_t} \mu_t)$$

where $F'_\mu \in \mathcal{C}^1(\mathbb{R}^p)$ is the Fréchet derivative of F at μ.

[Refs]:

Nitanda, Suzuki (2017). *Stochastic particle gradient descent for infinite ensembles.*

Mei, Montanari, Nguyen (2018). *A Mean Field View of the Landscape of Two-Layers Neural Networks.*

Rotskoff, Vanden-Eijndem (2018). *Parameters as Interacting Particles [...].*

Sirignano, Spiliopoulos (2018). *Mean Field Analysis of Neural Networks.*

Chizat, Bach (2018). *On the Global Convergence of Gradient Descent for Over-parameterized Models [...]*

Global convergence (Chizat & Bach 2018)

Theorem (2-homogeneous case)

Assume that ϕ is positively 2-homogeneous and some regularity. If the support of μ_0 covers all directions (e.g. Gaussian) and if $\mu_t \to \mu_\infty$ in $\mathcal{P}_2(\mathbb{R}^p)$, then μ_∞ is a global minimizer of F.

$\rightsquigarrow$ Non-convex landscape : initialization matters

Global convergence (Chizat & Bach 2018)

Theorem (2-homogeneous case)

Assume that ϕ is positively 2-homogeneous and some regularity. If the support of μ_0 covers all directions (e.g. Gaussian) and if $\mu_t \to \mu_\infty$ in $\mathcal{P}_2(\mathbb{R}^p)$, then μ_∞ is a global minimizer of F.

$\rightsquigarrow$ Non-convex landscape : initialization matters

Corollary

Under the same assumptions, if at initialization $\mu_{0,m} \to \mu_0$ then

$$\lim_{t\to\infty} \lim_{m\to\infty} F(\mu_{m,t}) = \lim_{m\to\infty} \lim_{t\to\infty} F(\mu_{m,t}) = \inf F.$$

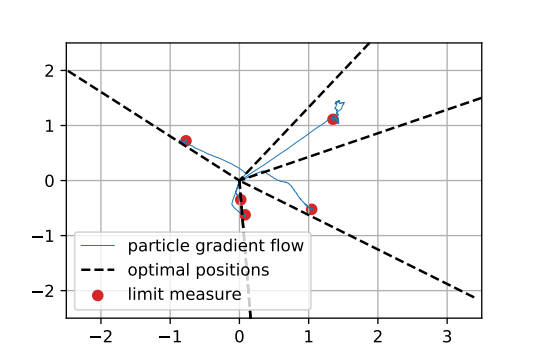
Generalization properties, if F is ...

- the **regularized empirical risk**: statistical adaptivity !
- the **population risk**: need convergence speed (?)
- the **unregularized empirical risk**: need implicit bias (?)

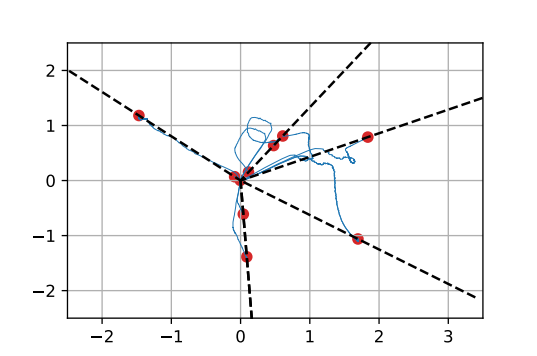
[Refs]:
Chizat, Bach (2018). *On the Global Convergence of Gradient Descent for Over-parameterized Models [...].*

Numerical Illustrations

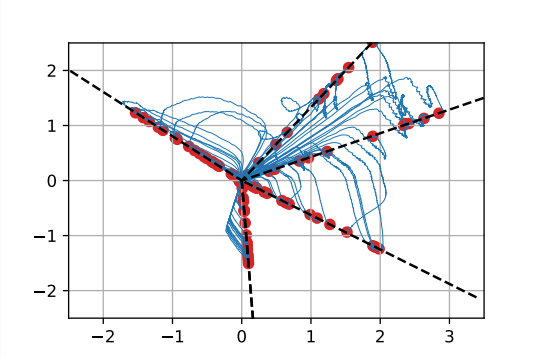
ReLU, $d = 2$, optimal predictor has 5 neurons (population risk)



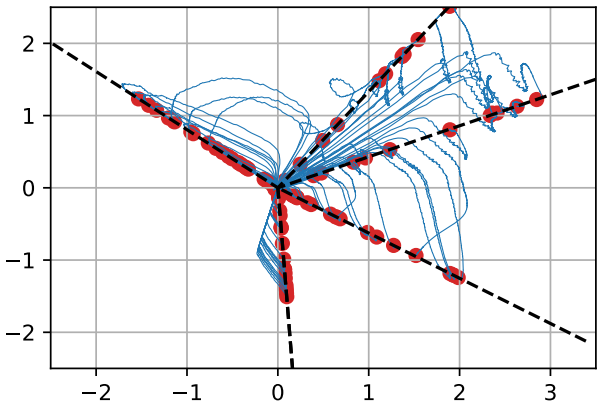
5 neurons



10 neurons



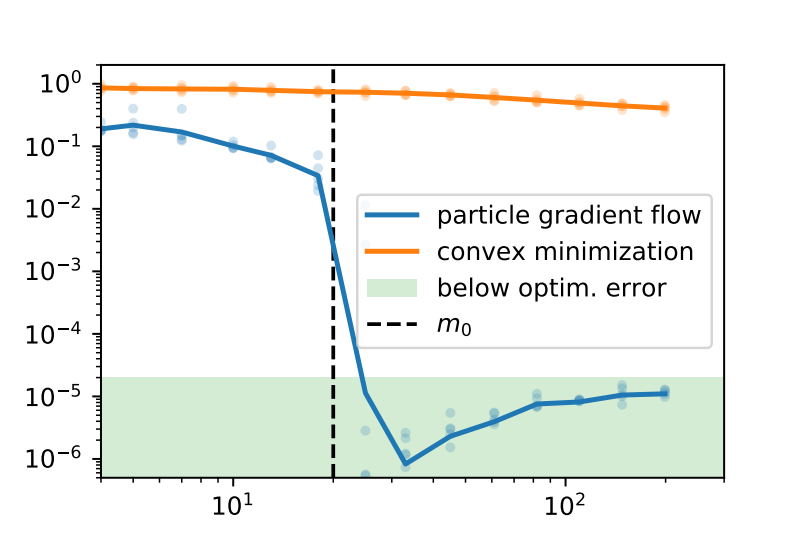
100 neurons



Performance

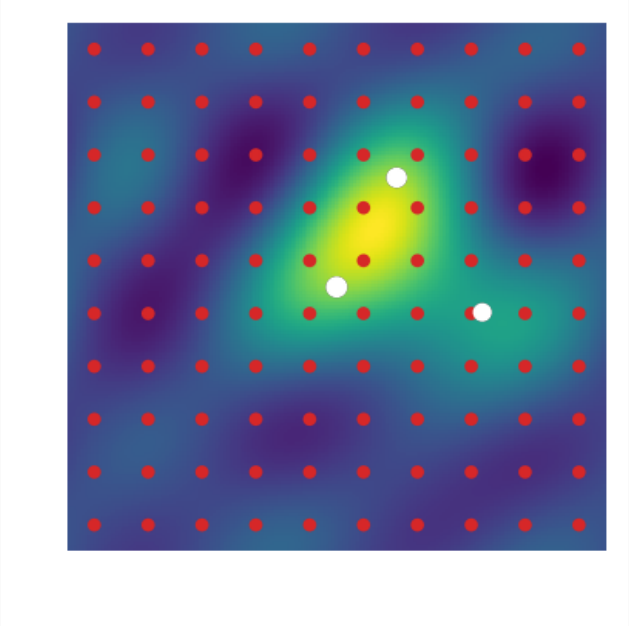
Population risk at convergence vs m

ReLU, $d = 100$, optimal predictor has 20 neurons



Convex optimization on measures

Sparse deconvolution on $\mathbb{T}^2$ with Dirichlet kernel
(white) sources (red) particles.



$\rightsquigarrow$ Computational guaranties for the regularized case, but m exponential in the dimension d

[Refs]:
Chizat (2019). *Sparse Optimization on Measures with Over-parameterized Gradient Descent.*

Lazy Training

Neural Tangent Kernel (Jacot et al. 2018)

Infinite width limit of standard neural networks

For infinitely wide fully connected neural networks of any depth with "standard" initialization and no regularization: the gradient flow implicitly performs kernel ridge(less) regression with the *neural tangent kernel*

$$K(x, x') = \lim_{m \to \infty} \langle \nabla_w \tilde{f}_m(w_0, x), \nabla_w \tilde{f}_m(w_0, x') \rangle.$$

Neural Tangent Kernel (Jacot et al. 2018)

Infinite width limit of standard neural networks

For infinitely wide fully connected neural networks of any depth with "standard" initialization and no regularization: the gradient flow implicitly performs kernel ridge(less) regression with the *neural tangent kernel*

$$K(x, x') = \lim_{m \to \infty} \langle \nabla_w \tilde{f}_m(w_0, x), \nabla_w \tilde{f}_m(w_0, x') \rangle.$$

Reconciling the two views:

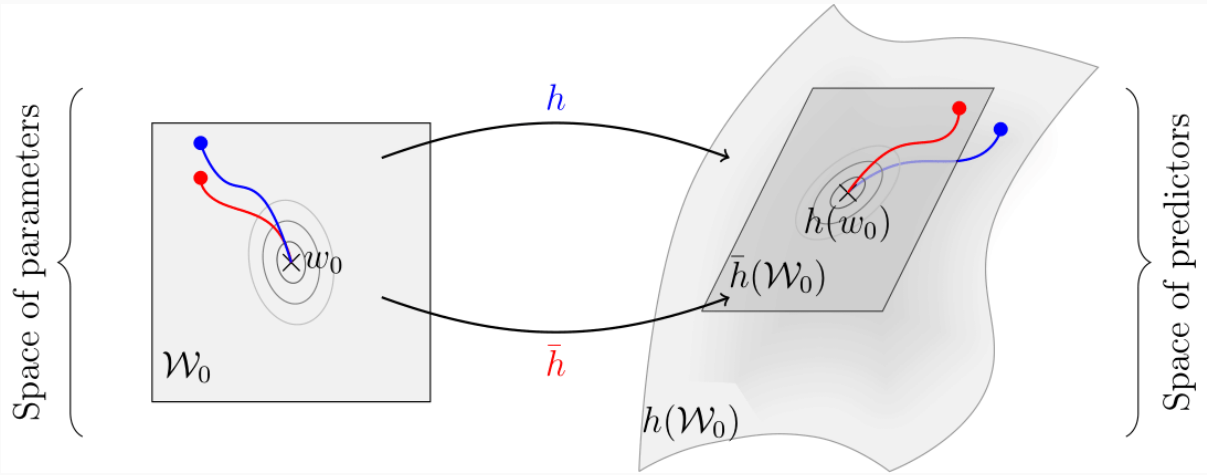
$$\tilde{f}_m(w, x) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \phi(w_i, x) \quad \textit{vs.} \quad f_m(w, x) = \frac{1}{m} \sum_{i=1}^{m} \phi(w_i, x)$$

This behavior is not intrinsically due to over-parameterization but to an exploding scale

[Refs]:
Jacot, Gabriel, Hongler (2018). *Neural Tangent Kernel: Convergence and Generalization in Neural Networks.*

Linearized model and scale

- let $h(w) = f(w, \cdot)$ be a differentiable model
- let $\bar{h}(w) = h(w_0) + Dh_{w_0}(w - w_0)$ be its linearization at w_0



Compare 2 training trajectories starting from w_0, with scale $\alpha > 0$:

- $w_\alpha(t)$ gradient flow of $F_\alpha(w) = R(\alpha h(w))/\alpha^2$
- $\bar{w}_\alpha(t)$ gradient flow of $\bar{F}_\alpha(w) = R(\alpha \bar{h}(w))/\alpha^2$

$\rightsquigarrow$ if $h(w_0) \approx 0$ and α large, then $w_\alpha(t) \approx \bar{w}_\alpha(t)$

Lazy training theorems

Theorem (Non-asymptotic)

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

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

Theorem (Strongly convex)

If $h(w_0) = 0$, and R strongly convex, it holds

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

- instance of *implicit bias*: *lazy* because parameters hardly move
- may replace the model by its linearization

[Refs]:

Chizat, Oyallon, Bach (2018). *On Lazy Training in Differentiable Programming*.

When does lazy training occur (without α)?

Relative scale criterion

For $R(y) = \frac{1}{2}\|y - y^\star\|^2$, relative error at normalized time t is

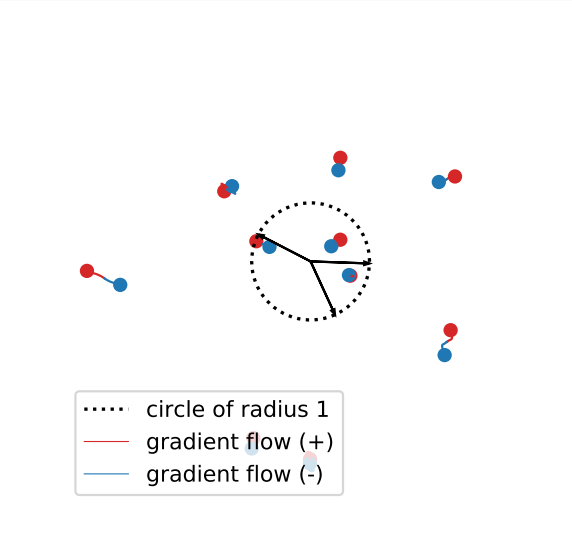
$$\text{err} \lesssim t^2 \kappa_h(w_0) \quad \text{where} \quad \kappa_h(w_0) := \frac{\|h(w_0) - y^\star\|}{\|\nabla h(w_0)\|} \frac{\|\nabla^2 h(w_0)\|}{\|\nabla h(w_0)\|}$$

Examples ($h(w) = f(w, \cdot)$):

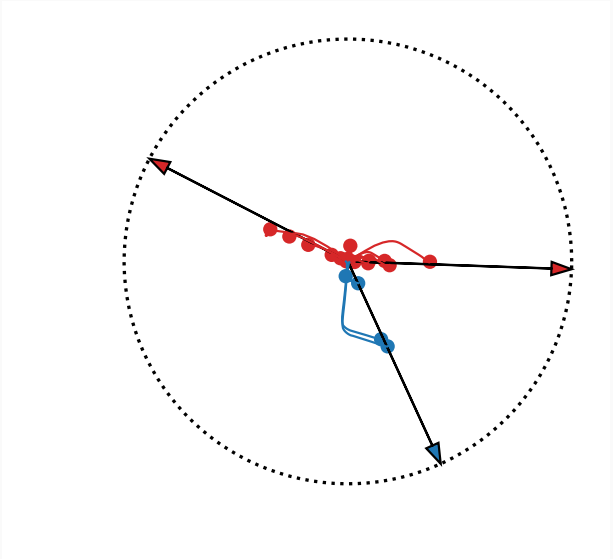
- *Homogeneous models with $f(w_0, \cdot) = 0$.*
  If for $\lambda > 0$, $f(\lambda w, x) = \lambda^L f(w, x)$, then $\kappa_f(w_0) \asymp 1/\|w_0\|^L$
- *Wide two-layer NNs with iid weights, $\mathbb{E}\Phi(w_i, \cdot) = 0$.*
  If $f(w, x) = \alpha(m) \sum_{i=1}^m \Phi(w_i, x)$, then $\kappa_f(w_0) \asymp (m\alpha(m))^{-1}$

Numerical Illustrations

Training paths (ReLU, $d = 2$, optimal predictor has $m = 3$ neurons)



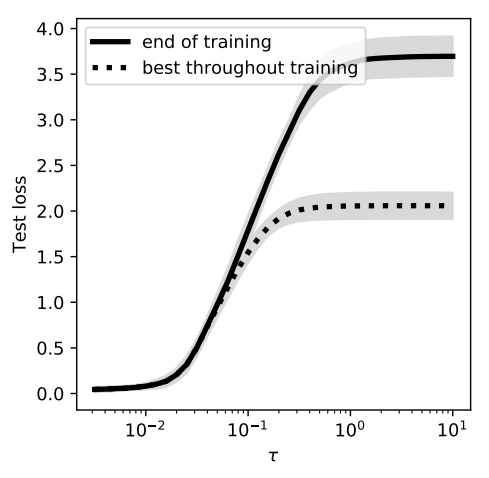
(a) Lazy



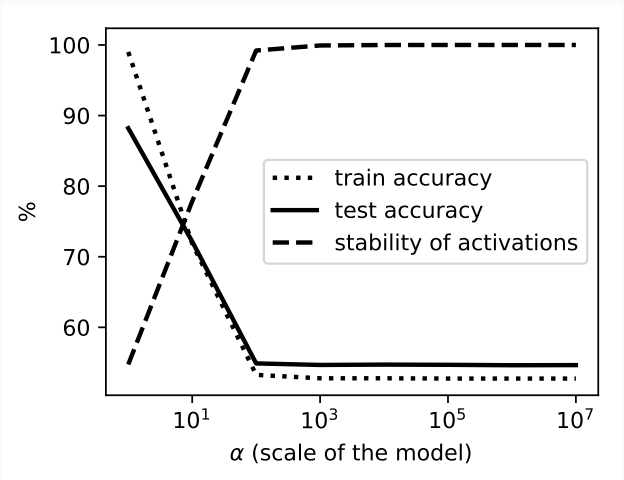
(b) Not lazy

Performance

Teacher-student setting, generalization in 100-d vs init. scale τ:



Perf. of ConvNets (VGG11) for image classification (CIFAR10):



- similar gaps observed for widened ConvNets & ResNets
- CKN[a] and taylored NTK[b] perform well on this (not so hard) task

[a] Convolutional Kernel Networks
[b] Neural Tangent Kernel

Conclusion

- Gradient descent on infinitely wide 2-layer networks converges to global minimizers
- Generalization behavior depends on initialization, loss, stopping time, signal scale, regularization...
- For the regularized empirical risk, it breaks the *statistical* curse of dimensionality
- But not (yet) the *computational* curse of dimensionality

[Refs]:
- Chizat, Bach (2018). *On the Global Convergence of Over-parameterized Models using Optimal Transport.*
- Chizat, Oyallon, Bach (2019). *On Lazy Training in Differentiable Programming.*
- Chizat (2019). *Sparse Optimization on Measures with Over-parameterized Gradient Descent.*