Analyses of gradient methods for the optimization of wide two-layer neural networks

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

Jan. 9, 2020 - Statistical Physics and Machine Learning - ICTS

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

Introduction

Setting

Supervised machine learning

- given input/output training data $(x^{(1)}, y^{(1)}), \dots, (x^{(n)}, y^{(n)})$
- 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 (learning in a Hilbert space)

Linear regression, prior/random features, kernel methods:

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

Models

Linear in the parameters (learning in a Hilbert space)

Linear regression, prior/random features, kernel methods:

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

Neural networks

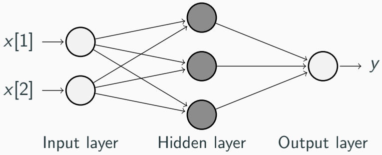
Vanilla NN with activation σ & parameters $(W_1, b_1), \ldots, (W_L, b_L)$:

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

Content

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

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

- Empirical risk: $\frac{1}{n} \sum_i \ell(f(x_i), y_i)$, population risk: $\mathbb{E}[\ell(f(x), y)]$

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

$$\min_{\mu \in \mathcal{P}(\mathbb{R}^p)} F(\mu) := \underbrace{R(f(\mu, \cdot))}_{\text{Empirical or population risk}} + \underbrace{\lambda G(\mu)}_{\text{Regularization}}$$

- lifted version of "convex" neural networks
- in next slide: with $V(w) = \|a\|^2 + |b|^2 + |c|^2$, let $G(\mu) = \int V d\mu$ and solve (for well chosen δ depending on n):

$$\min_{\mu \in \mathcal{P}(\mathbb{R}^p)} \frac{1}{n} \sum_{i=1}^{n} |f(\mu, x_i) - y_i|^2 \quad \text{subject to} \quad G(\mu) \leq \delta$$

[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)

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

- $\tilde{O}(n^{-1/d})$ 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 a properly tuned regularization level

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

$\rightsquigarrow$ 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.*

Neural networks and random features

What happens when training only the output layer?

- fix a distribution $\mathrm{d}\tau$ of the hidden layer weights
- this leads to the parametric model

$$f(g, x) = \int \sigma(a \cdot x + b) g(a, b) \mathrm{d}\tau(a, b) \quad \text{for} \quad g \in L^2(\mathrm{d}\tau)$$

- training the output layer, i.e. solving

$$\min_{g \in L^2(\mathrm{d}\tau)} R(f(g, \cdot)) + \lambda \|g\|^2_{L^2(\mathrm{d}\tau)}$$

  amounts to kernel ridge regression
- the *conjugate* kernel is given by

$$k(x, x') = \int \sigma(a \cdot x + b) \sigma(a \cdot x' + b) \mathrm{d}\tau(a, b)$$

$\rightsquigarrow$ later we will also see the *tangent* kernel

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 $w_1(0), w_2(0), \dots$ *are such that* $\mu_{0,m} \to \mu_0$ *in* $\mathcal{P}_2(\mathbb{R}^p)$ *and technical assumptions. 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 [...]*

Parenthesis (I): square loss and interacting particles system

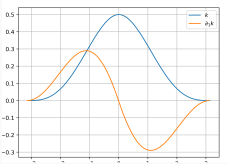
When R is the square loss w.r.t input density ρ:

$$F(\mu) = \frac{1}{2}\int \left\| \int \phi(w,x)\mathrm{d}\mu(w) - \int \phi(w,x)\mathrm{d}\mu^\star(w) \right\|^2 \mathrm{d}\rho(x)$$

$$= \frac{1}{2}\int\int k(w,w')\mathrm{d}\mu(x)\mathrm{d}\mu(w') - \int\int k(w,w')\mathrm{d}\mu(x)\mathrm{d}\mu^\star(w') + C$$

where $k(w,w') = \int \phi(w,x)\phi(w',x)\mathrm{d}\rho(x)$.

- this is the mean-field description of a system of interacting particle with interaction potential $k(w,w')$
- adding noise $\sim$ adding an entropy term in F



k as a function of angle(w,w') for Relu and uniform ρ

Parenthesis (II): homogeneity

In preparation of the global convergence result, we need to define:

2-homogeneity

We say that $\phi(w, x)$ is positively 2-homogeneous if

$$\phi(\lambda w, x) = \lambda^2 \phi(w, x), \quad \forall \lambda > 0$$

2-homogeneous projection

For a measure $\mu \in \mathcal{P}_2(\mathbb{R}^p)$, its projection $\Pi_2(\mu) \in \mathcal{M}_+(\mathbb{S}^{p-1})$ is characterized by the property that $\forall \varphi \in \mathcal{C}(\mathbb{S}^{p-1})$,

$$\int_{\mathbb{S}^{p-1}} \varphi(w) \mathrm{d}[\Pi_2(\mu)](w) = \int_{\mathbb{R}^p} \|w\|_2^2 \varphi(w/\|w\|_2) \mathrm{d}\mu(w)$$

Global convergence (C. & Bach 2018)

Theorem (2-homogeneous case)

Assume ϕ is positively 2-homogeneous and technical assumptions. If $\Pi_2(\mu_0)$ is full support on $\mathbb{S}^{p-1}$ (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 (C. & Bach 2018)

Theorem (2-homogeneous case)

Assume ϕ is positively 2-homogeneous and technical assumptions. If $\Pi_2(\mu_0)$ is full support on $\mathbb{S}^{p-1}$ (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

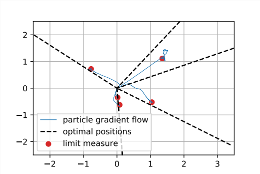
$$\lim_{t\to\infty} \lim_{m\to\infty} F(\mu_{m,t}) = \lim_{m\to\infty} \lim_{t\to\infty} F(\mu_{m,t}) = \inf F.$$

Generalization properties, if F is ...

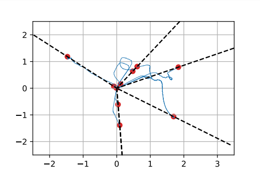
- the **regularized empirical risk**: statistical adaptivity !
- the **population risk**: need convergence speed (?)
- the **unregularized empirical risk**: need implicit bias (?)

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

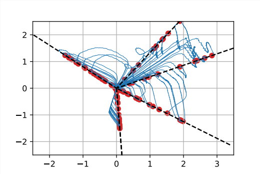
ReLU, $d = 2$, optimal predictor has 5 neurons (population risk)



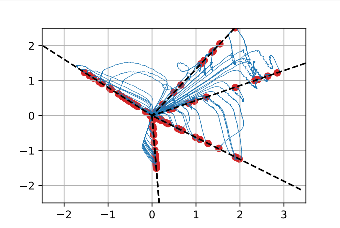
5 neurons



10 neurons



100 neurons

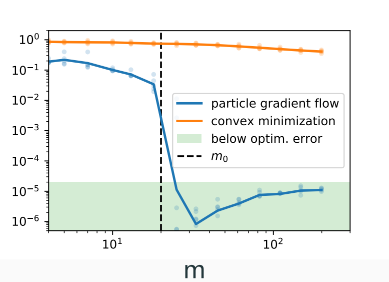


Proof idea:
(i) given a suboptimal stationary point, the dynamics escapes its neighborhoods if a certain set of directions contains a particle
(ii) such a particle always exists with full support initialization

Empirical performance

Population risk at convergence vs m

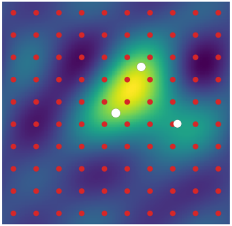
ReLU, $d = 100$, optimal predictor has 20 neurons



Optimization on measures and quantitative results

Beyond neural networks

- ideas relevant for various optimizations problems on measures
- computational guaranties and exponential local convergence, **but:**
  - for the regularized case with a non-degeneracy condition
  - requires m exponential in the dimension d
  - forces the mass of particles to vary faster than their position



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

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

Optimization on Measures

Setting

- Θ compact d-Riemannian manifold without boundaries
- $\mathcal{M}_+(\Theta)$ nonnegative finite Borel measures
- $\phi : \Theta \to \mathcal{F}$ smooth, $\mathcal{F}$ separable Hilbert space
- $R : \mathcal{F} \to \mathbb{R}_+$ convex and smooth, $\lambda > 0$

$$\min_{\nu \in \mathcal{M}_+(\Theta)} J(\nu) := R\left(\int_\Theta \phi(\theta)\mathrm{d}\nu(\theta)\right) + \lambda\nu(\Theta)$$

In this section

Simple non-convex gradient descent algorithms reaching ϵ-accuracy in $O(\log(1/\epsilon))$ complexity under non-degeneracy assumptions.

$\rightsquigarrow$ the signed case can be covered by "doubling" the space $\rightsquigarrow$ continuous, infinite dimensional LASSO problem

Particle Gradient Descent

Algorithm (general case)

- initialize with discrete measure $\nu = \frac{1}{m}\sum_{i=1}^{m} r_i^p \delta_{\theta_i}$, with $p \geq 1$
- run gradient descent (or variant) on $(r_i, \theta_i)^m \in (\mathbb{R}_+ \times \Theta)^m$

Questions for theory

1. What choice for p? for the metric on $\mathbb{R}_+ \times \Theta$?
2. Is it a consistent method? for which initialization?
3. Are there computational complexity guarantees?

Conic Particle Gradient Descent

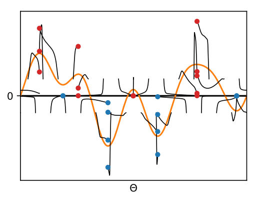
Algorithm (conic particle gradient descent)

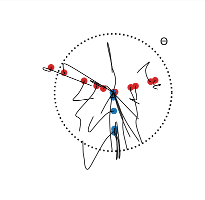
Take $p = 2$ and discretize (with retractions) the gradient flow

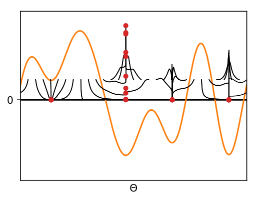
$$\begin{cases} r_i'(t) = -2\alpha r_i(t) J'_{\nu_t}(\theta_i(t)) \\ \theta_i'(t) = -\beta \nabla J'_{\nu_t}(\theta_i(t)). \end{cases}$$

where $J'_\nu(\theta) = \langle \phi(\theta), \nabla R(\int \phi \mathrm{d}\nu)\rangle + \lambda$ “is” the Fréchet derivative of J at ν and $\nu_t = \frac{1}{m}\sum_{i=1}^m r_i(t)^2 \delta_{\theta_i(t)}$.

$\rightsquigarrow$ gradient flow in the *Wasserstein-Fisher-Rao metric* of $\mathcal{M}_+(\Theta)$







Sharpness/Polyak-Łojasiewicz Inequality

Theorem (C., 2019)

Under the non-degenerate dual certificate assumption, $\exists J_0, \kappa_0 > 0$ such that for any $\nu \in \mathcal{M}_+(\Theta)$ satisfying $J(\nu) \leq J_0$, it holds

$$\underbrace{\int \left(4\alpha|J'_\nu|^2 + \beta\|\nabla J'_\nu\|^2_\theta\right) \mathrm{d}\nu}_{\textit{Squared-norm of gradient}} \geq \kappa_0 \min\{\alpha, \beta\} \underbrace{(J(\nu) - J^\star)}_{\textit{Optimality gap}}.$$

- J_0 and κ_0 polynomial in the problem characteristics
- crucially, J_0 is independent of the over-parameterization

Corollary

If $J(\nu_0) \leq J_0$, gradient flow and gradient descent (with $\max\{\alpha, \beta\}$ small enough) converge exponentially fast to the global minimizer, in value and in distance (e.g. Bounded-Lipschitz, WFR).

Quantitative Global Convergence

Fine tuning

Particule gradient descent can be used after any optimization algorithm : discrete convex optimization, conditional gradient...

$\rightsquigarrow$ Can we use a single algorithm?

Quantitative Global Convergence

Fine tuning

Particule gradient descent can be used after any optimization algorithm : discrete convex optimization, conditional gradient...

$\leadsto$ Can we use a single algorithm?

Theorem (C., 2019)

For $M, \epsilon > 0$ fixed, there exists $C_1, C_2 > 0$ such that if for $\eta > 0$,

$$W_\infty(\nu_0, M\mathrm{vol}) < C_1\eta \qquad \textit{and} \qquad \frac{\beta}{\alpha} < C_2\eta^{2(1+\epsilon)}$$

then for $\alpha t \geq C_3/\eta^{1+\epsilon}$ it holds $J(\nu_t) - J^\star \leq \eta$. It particular, if this holds for $\eta = J_0 - J^\star$, then $(\nu_t)_{t\geq 0}$ converges to a global minimizer.

Via samples or a grid: $W_\infty(\nu_0, \mathrm{vol}) \asymp m^{-1/d}$ with m particles.

$\leadsto$ not suitable for high dimensional machine learning problems

Proof Idea: Perturbed Mirror Descent

Lemma (Mirror descent rate)

The dynamic with $\beta = 0$ satisfies, for some $C > 0$,

$$J(\nu_t) - J(\nu^\star) \lesssim \inf_{\nu \in \mathcal{M}_+(\Theta)} \left\{ \|\nu^\star - \nu\|_{BL} + \frac{1}{Ct} \mathcal{H}(\nu, \nu_0) \right\}$$
$$\lesssim \frac{\log t}{t} \qquad \text{if } \nu_0 \propto \mathrm{d}\,\mathrm{vol}$$

where $\mathcal{H}$ is the relative entropy (Kullback-Leibler divergence).

Proof idea of the theorem.

Adapt this lemma to deal with $\beta > 0$, to show that the dynamics reaches the sublevel of exponential convergence. □

- discrete time dynamics & choice of retraction: see paper

Exponential loss and implicit bias

Training linear models with exponential loss

Let $(x_i, y_i)_{i=1}^n$ with $y_i \in \{-1, +1\}$, no regularization & exponential loss

$$R(f) = \frac{1}{n} \sum_{i=1}^{n} \exp(-y_i f(x_i))$$

Theorem (Soudry et al. 2018, simplified)

Consider $f(w, x) = w^\intercal x$ and a linearly separable data set. For any initialization, the normalized gradient flow $\bar{w}(t) = w(t)/\|w(t)\|_2$ converges to a max-margin classifier, solution to

$$\max_{\|w\|_2 \leq 1} \min_{i \in [n]} \; y_i \cdot w^\intercal x_i$$

- also applies to the logistic loss (a.k.a. cross-entropy)
- many extensions; but no global result for non-convex models

[Refs]:
Soudry *et al.* (2018). *The Implicit Bias of Gradient Descent on Separable Data.*

An intuition behind the result

- look at $w'(t) = \nabla F_1(w(t))$, with the soft-min loss

$$F_\beta(w) = -\frac{1}{\beta} \log \left(\frac{1}{n} \sum_{i=1}^{n} \exp(-\beta y_i w^\intercal x_i) \right) \xrightarrow[\beta \to \infty]{} \min_i y_i w^\intercal x_i$$

- observe that $\|w(t)\| \to \infty$ for separable data sets
- denoting $\bar{w}(t) = w(t)/\|w(t)\|_2$, it holds

$$\frac{d}{dt}\bar{w}(t) = \frac{1}{\|w(t)\|} \nabla F_{\|w(t)\|}(\bar{w}(t)) - \alpha_t \bar{w}(t)$$

for some $\alpha_t > 0$ that constraints $\bar{w}(t)$ to the sphere

- thus $\bar{w}(t)$ performs online projected gradient ascent

Implicit bias of two-layer neural networks

Consider the mean-field dynamic $\mu_t \in \mathcal{P}_2(\mathbb{R}^p)$ for the exponential loss, and its projection on the sphere $\nu_t = \Pi_2(\mu_t)$, satisfying

$$\int \varphi(\theta)\mathrm{d}\nu_t(\theta) = \int |u|^2 \varphi(u/|u|)\mathrm{d}\mu_t(u) \qquad \forall \varphi \in \mathcal{C}(\mathbb{S}^{p-1})$$

Theorem (C. and Bach)

Assume that ϕ is 2-homogeneous and technical assumptions. If:

- $y_i \neq y_j \quad \Rightarrow \quad x_i \neq x_j$ *(for non-polynomial activation)*
- ν_0 *has full support on the sphere, and*
- ν_t *and* $\nabla R(f(\nu_t, \cdot))$ *converge in direction*

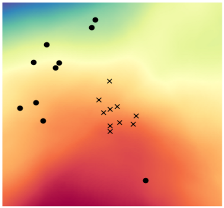
then $\lim_{t\to\infty} \nu_t / \|\nu_t\|_{TV}$ *solves*

$$\max_{\|\nu\|_{TV} \leq 1} \min_i \ y_i f(\nu, x_i)$$

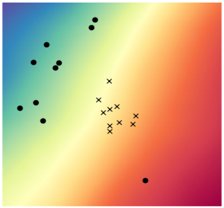
[Refs]:

Chizat, Bach (in preparation). *Implicit bias of wide two-layer neural networks trained for the exponential loss.*

Illustration



Two-layer ReLU neural net



Random hidden layer fixed

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*

$$\langle \nabla_w \tilde{f}_m(w_0, x), \nabla_w \tilde{f}_m(w_0, x') \rangle \xrightarrow[m \to \infty]{p} K(x, x').$$

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*

$$\langle \nabla_w \tilde{f}_m(w_0, x), \nabla_w \tilde{f}_m(w_0, x') \rangle \xrightarrow[m\to\infty]{p} K(x, x').$$

Reconciling the two views:

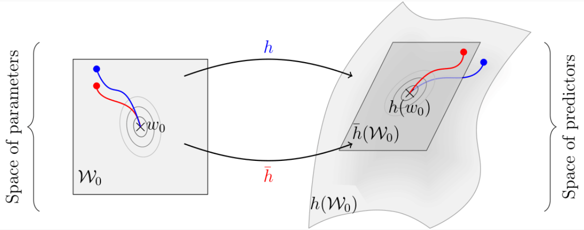
$$\tilde{f}_m(w, x) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \phi(w_i, x) \quad vs. \quad f_m(w, x) = \frac{1}{m} \sum_{i=1}^{m} \phi(w_i, x)$$

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

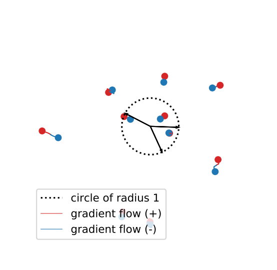
$$\text{err} \lesssim t^2 \kappa_h(w_0) \quad \text{where} \quad \kappa_h(w_0) := \frac{\|h(w_0) - y^\star\|}{\|\nabla h(w_0)\|} \frac{\|\nabla^2 h(w_0)\|}{\|\nabla h(w_0)\|}$$

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

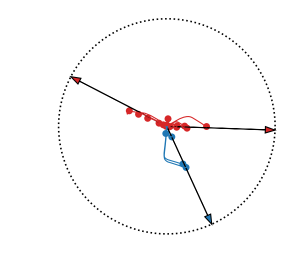
- *Homogeneous models with $f(w_0, \cdot) = 0$.*
  If for $\lambda > 0$, $f(\lambda w, x) = \lambda^L f(w, x)$, then $\kappa_f(w_0) \asymp 1/\|w_0\|^L$
- *Wide two-layer NNs with iid weights, $\mathbb{E}\Phi(w_i, \cdot) = 0$.*
  If $f(w, x) = \alpha(m) \sum_{i=1}^m \Phi(w_i, x)$, then $\kappa_f(w_0)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

Numerical Illustrations

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



(a) Lazy



(b) Not lazy

- for linear classifiers, implicit bias characterized for $\alpha \in (0, \infty)$ in (Woodworth et al., 2019)

[See also]:
Woodworth, Gunasekar, Lee, Soudry, Srebro (2019). *Kernel and Deep Regimes in Overparametrized Models.*

Performance

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

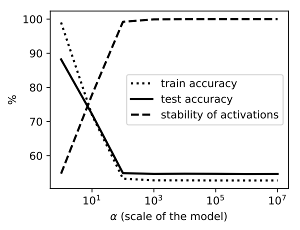


Figure 5: VGG-11 on CIFAR10

- similar gaps observed for widened ConvNets & ResNets
- in depth exploration in Mario's talk tomorrow

Conclusion

- Gradient descent on infinitely wide 2-layer networks converges to global minimizers
- Generalization behavior depends on initialization, loss, stopping time, signal scale, regularization...
- open questions: quantitative results, more layers

[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.*
- Chizat, Bach (in preparation). *Implicit bias of wide two-layer neural networks trained with the exponential loss.*