



Analysis of Gradient Descent on Wide Two-Layer ReLU Neural Networks

Lénaïc Chizat^{*}, joint work with Francis Bach⁺

Oct 22nd 2021 - FAU Erlangen

^{*}EPFL ⁺INRIA and ENS Paris

Supervised learning with neural networks

Prediction/classification task

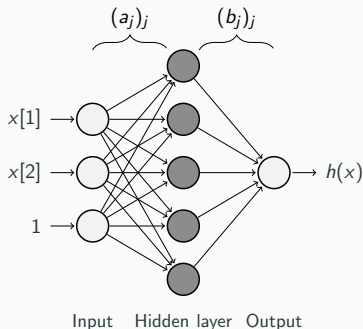
- Couple of random variables (X, Y) on $\mathbb{R}^d \times \mathbb{R}$
- Given n i.i.d. samples $(x_i, y_i)_{i=1}^n$, build h s.t. $h(X) \approx Y$

Wide 2-layer ReLU neural network

For a width $m \gg 1$, predictor h given by

$$h((w_j)_j, x) := \frac{1}{m} \sum_{j=1}^m \phi(w_j, x)$$

where $\begin{cases} \phi(w, x) := b(a^\top [x; 1])_+ \\ w := (a, b) \in \mathbb{R}^{d+1} \times \mathbb{R} \end{cases}$



$\rightsquigarrow \phi$ is 2-homogeneous in w , i.e. $\phi(rw, x) = r^2 \phi(w, x), \forall r > 0$

Gradient flow of the empirical risk

$$\text{Convex smooth loss } \ell: \begin{cases} \ell(p, y) = \log(1 + \exp(-yp)) & \text{(logistic)} \\ \ell(p, y) = (y - p)^2 & \text{(square)} \end{cases}$$

Empirical risk with weight decay ($\lambda \geq 0$)

$$F_m((w_j)_j) := \underbrace{\frac{1}{n} \sum_{i=1}^n \ell(h((w_j)_j, x_i), y_i)}_{\text{empirical risk}} + \underbrace{\frac{\lambda}{m} \sum_{j=1}^m \|w_j\|_2^2}_{\text{(optional) regularization}}$$

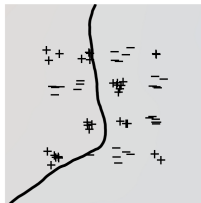
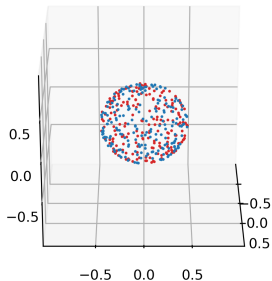
Gradient flow

- Initialize $w_1(0), \dots, w_m(0) \stackrel{\text{i.i.d}}{\sim} \mu_0 \in \mathcal{P}_2(\mathbb{R}^{d+1} \times \mathbb{R})$
- Decrease the non-convex objective via gradient flow, for $t \geq 0$,

$$\frac{d}{dt}(w_j(t))_j = -m \nabla F_m((w_j(t))_j)$$

\rightsquigarrow in practice, discretized with variants of gradient descent

Illustration : logistic loss, unregularized ($\lambda = 0$)



Space of parameters

- plot $|b_j| \cdot a_j$
- color depends on sign of b_j
- tanh radial scale

Space of predictors

- (+/-) training set
- color shows $h((w_j(t)))_j, \cdot)$
- line shows 0 level set

Main question

What is performance of the learnt predictor $h((w_j(\infty)))_j, \cdot$?

Infinite width limit: global convergence

Regularized case: function spaces

Unregularized case: implicit regularization

Infinite width limit: global convergence

Dynamics in the infinite width limit

- Parameterize with a probability measure $\mu \in \mathcal{P}_2(\mathbb{R}^{d+2})$

$$h(\mu, x) = \int \phi(w, x) d\mu(w)$$

- Objective on the space of probability measures

$$F(\mu) := \frac{1}{n} \sum_{i=1}^n \ell(h(\mu, x_i), y_i) + \lambda \int \|w\|_2^2 d\mu(w)$$

Theorem (dynamical infinite width limit, adapted to ReLU)

Assume that

$$\text{spt}(\mu_0) \subset \{(a, b) \in \mathbb{R}^{d+1} \times \mathbb{R} ; \|a\|_2 = |b|\}.$$

As $m \rightarrow \infty$, $\mu_{t,m} = \frac{1}{m} \sum_{j=1}^m \delta_{w_j(t)}$ converges a.s. in $\mathcal{P}_2(\mathbb{R}^{d+2})$ to μ_t , the unique Wasserstein gradient flow of F starting from μ_0 .

Theorem (C. & Bach, '18, adapted to ReLU)

Assume that $\mu_0 = \mathcal{U}_{\mathbb{S}^d} \otimes \mathcal{U}_{\{-1,1\}}$ and technical conditions. If μ_t converges weakly to μ_∞ , then μ_∞ is a global minimizer of F .

- Initialization matters: the key assumption on μ_0 is *diversity*
- Corollary: $\lim_{m,t \rightarrow \infty} F(\mu_{m,t}) = \min F$
- Open question: convergence of μ_t

Performance of the learnt predictor?

Depends on the objective F and the data! If F is the ...

- **regularized empirical risk**: “just” statistics (this talk)
- **unregularized empirical risk**: need implicit bias (this talk)
- **population risk**: need convergence speed (open question)

Illustration: teacher student

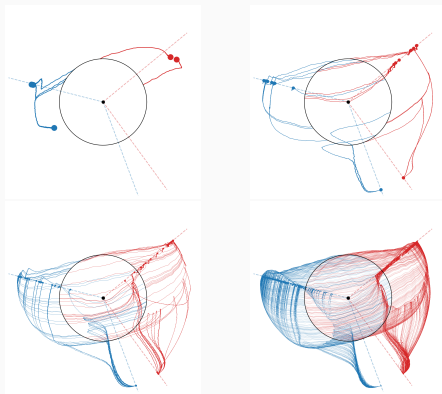


Figure 1: SGD on expected square loss,

$$X \sim \mathcal{U}_{\mathbb{S}^d} \text{ and } Y = h((w_i^*)_{i=1}^{m^*}, X)$$

[Related work studying infinite width limits]:

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.*

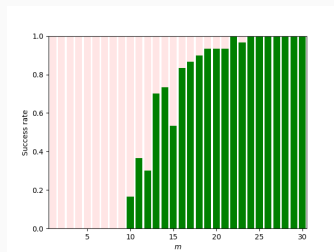


Figure 2: Success rate when
 $d = 100, m^* = 10$

Regularized case: function spaces

Definition (Variation norm)

For a predictor $h : \mathbb{R}^d \rightarrow \mathbb{R}$, its variation norm is

$$\begin{aligned}\|h\|_{\mathcal{F}_1} &:= \min_{\mu \in \mathcal{P}_2(\mathbb{R}^{d+2})} \left\{ \frac{1}{2} \int \|w\|_2^2 d\mu(w) ; h(x) = \int \phi(w, x) d\mu(w) \right\} \\ &= \min_{\nu \in \mathcal{M}(\mathbb{S}^d)} \left\{ \|\nu\|_{TV} ; h(x) = \int (a^\top [x; 1])_+ d\nu(a) \right\}\end{aligned}$$

Proposition

If $\mu^* \in \mathcal{P}_2(\mathbb{R}^{d+2})$ minimizes F then $h(\mu^*, \cdot)$ minimizes

$$\frac{1}{n} \sum_{i=1}^n \ell(h(x_i), y_i) + 2\lambda \|h\|_{\mathcal{F}_1}.$$

Fixing the hidden layer and conjugate RKHS

What if we only train the output layer?

\rightsquigarrow Let $\mathcal{S} := \{\mu \in \mathcal{P}_2(\mathbb{R}^{d+2}) \text{ with marginal } \mathcal{U}_{\mathbb{S}^d} \text{ on input weights}\}$

Definition (Conjugate RKHS)

For a predictor $h : \mathbb{R}^d \rightarrow \mathbb{R}$, its conjugate RKHS norm is

$$\|h\|_{\mathcal{F}_2}^2 := \min \left\{ \int |b|_2^2 d\mu(a, b) ; h = \int \phi(w, \cdot) d\mu(w), \mu \in \mathcal{S} \right\}$$

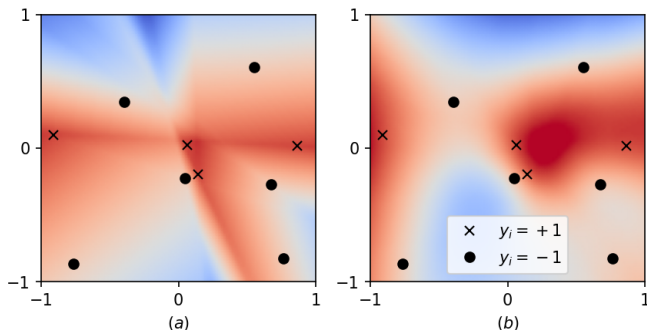
Proposition (Kernel ridge regression)

All else unchanged, fixing the hidden layer leads to minimizing

$$\frac{1}{n} \sum_{i=1}^n \ell(h(x_i), y_i) + \lambda \|h\|_{\mathcal{F}_2}^2.$$

Illustration of the predictor

Predictor learnt via gradient descent (square loss & weight decay)



(a) Training both layers (\mathcal{F}_1 -norm) (b) Training output layer (\mathcal{F}_2 -norm)

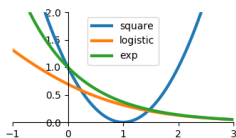
	\mathcal{F}_1	\mathcal{F}_2
Stat. prior	Adaptivity to anisotropy	Isotropic smoothness
Optim.	No guarantee	Guaranteed efficiency

Unregularized case: implicit regularization

Preliminary: linear classification with exponential loss

Classification task

- $Y \in \{-1, 1\}$ and prediction is $\text{sign}(h(X))$
- no regularization ($\lambda = 0$)
- loss with an exponential tail
 - exponential $\ell(p, y) = \exp(-py)$, or
 - logistic $\ell(p, y) = \log(1 + \exp(-py))$



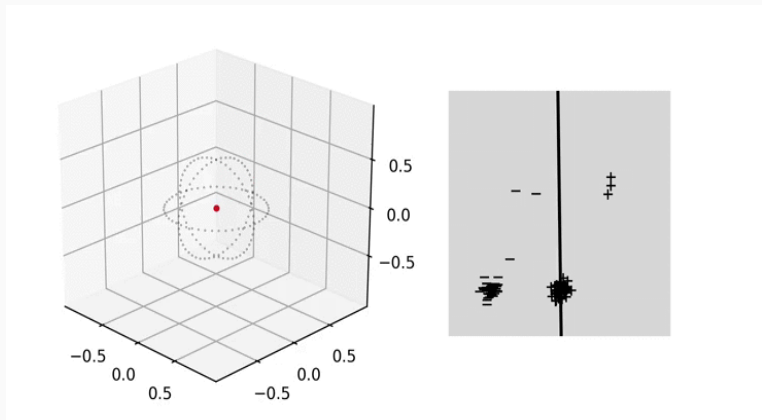
Loss for $y = 1$

Theorem (SHNGS 2018, reformulated)

Consider $h(w, x) = w^T x$ and a linearly separable training set. For any $w(0)$, the normalized gradient flow $\bar{w}(t) = w(t)/\|w(t)\|_2$ converges to a $\|\cdot\|_2$ -max-margin classifier, i.e. a solution to

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

Implicit regularization for linear classification: illustration



Implicit bias of gradient descent for classification ($d = 2$)

Implicit regularizations for 2-layer neural networks

Back to wide 2-layer ReLU neural networks.

Theorem (C. & Bach, 2020)

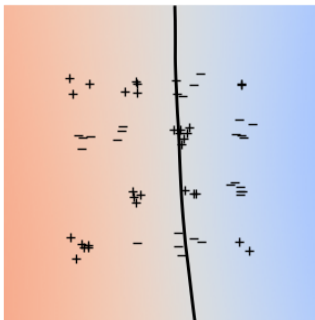
Assume that $\mu_0 = \mathcal{U}_{\mathbb{S}^d} \otimes \mathcal{U}_{\{-1,1\}}$, that the training set is consistent ($[x_i = x_j] \Rightarrow [y_i = y_j]$) and technical conditions (in particular, of convergence). Then $h(\mu_t, \cdot) / \|h(\mu_t, \cdot)\|_{\mathcal{F}_1}$ converges to the \mathcal{F}_1 -max-margin classifier, i.e. it solves

$$\max_{\|h\|_{\mathcal{F}_1} \leq 1} \min_{i \in [n]} y_i h(x_i).$$

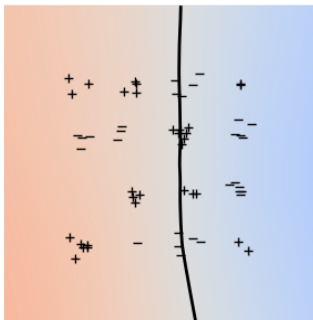
- fixing the hidden layer leads to the \mathcal{F}_2 -max-margin classifier
- we also prove convergence speed bounds in simpler settings

Illustration

Training output layer



Training both layers



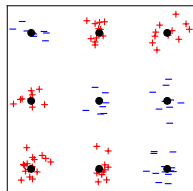
$h(\mu_t, \cdot)$ for the exponential loss, $\lambda = 0$ ($d = 2$)

Numerical experiments

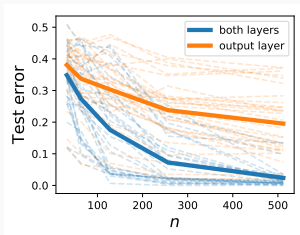
Setting

Two-class classification in dimension $d = 15$:

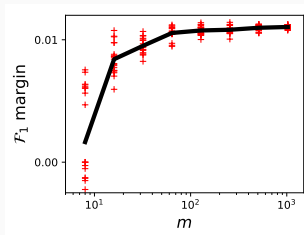
- two first coordinates as shown on the right
- all other coordinates uniformly at random



Coordinates 1 & 2



(a) Test error vs. n



(b) Margin vs. m ($n = 256$)

Statistical efficiency

Assume that $\|X\|_2 \leq D$ a.s. and that, for some $r \leq d$, it holds a.s.

$$\Delta(r) \leq \sup_{\pi} \left\{ \inf_{y_i \neq y_{i'}} \|\pi(x_i) - \pi(x_{i'})\|_2 ; \pi \text{ is a rank } r \text{ projection} \right\}.$$

Theorem (C. & Bach, 2020)

The \mathcal{F}_1 -max-margin classifier h^* admits the risk bound, with probability $1 - \delta$ (over the random training set),

$$\underbrace{\mathbf{P}(Y h^*(X) < 0)}_{\text{proportion of mistakes}} \lesssim \frac{1}{\sqrt{n}} \left[\left(\frac{D}{\Delta(r)} \right)^{r+2} + \sqrt{\log(1/\delta)} \right].$$

- this is a strong *dimension independent* non-asymptotic bound
- for learning in \mathcal{F}_2 the bound with $r = d$ is true
- this task is *asymptotically* easy (the rate $n^{-1/2}$ is suboptimal)

[Refs]:

Chizat, Bach (2020). *Implicit Bias of Gradient Descent for Wide Two-layer Neural Networks [...]*.

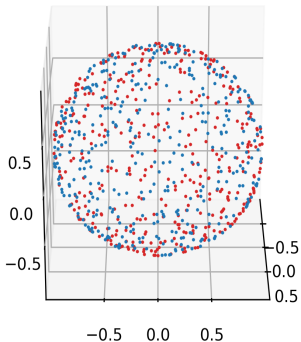
Two implicit regularizations in one dynamics (I)

Lazy training (informal)

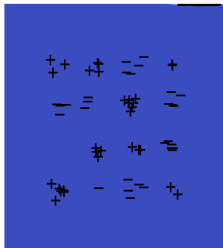
All other things equal, if the variance at initialization is large and the step-size is small then the model behaves like its first order expansion over a significant time.

- Neurons hardly move but significant total change in $h(\mu_t, \cdot)$
- Here, the linearization converges to a max-margin classifier in the tangent RKHS (similar to \mathcal{F}_2)
- Eventually converges to \mathcal{F}_1 -max-margin

Two implicit regularizations in one dynamics (II)



Space of parameters



Space of predictors

See also: Moroshko, Gunasekar, Woodworth, Lee, Srebro, Soudry (2020). *Implicit Bias in Deep Linear Classification: Initialization Scale vs Training Accuracy*.

- Open question: make statements of this talk quantitative
 - ↪ how fast is the convergence ? how many neurons are needed?
- Mathematical models for deeper networks
 - ↪ goal: formalize training dynamics & study generalization

[Talk based on the following papers:]

- Chizat, Bach (NeurIPS 2018). *On the Global Convergence of Over-parameterized Models using Optimal Transport.*
- Chizat, Oyallon, Bach (NeurIPS 2019). *On Lazy Training in Differentiable Programming.*
- Chizat, Bach (COLT 2020). *Implicit Bias of Gradient Descent for Wide Two-layer Neural Networks Trained with the Logistic Loss.*

Generalization with variation norm regularization

Regression of a Lipschitz function

Assume that X is bounded and $Y = f^*(X)$ where f^* is 1-Lipschitz. Error bound on $\mathbf{E}[(h(X) - f^*(X))^2]$ for any estimator h ?

\rightsquigarrow in general $\succeq n^{-1/d}$ unavoidable (curse of dimensionality)

Anisotropy assumption:

What if moreover $f^*(x) = g(\pi_r(x))$ for some rank r projection π_r ?

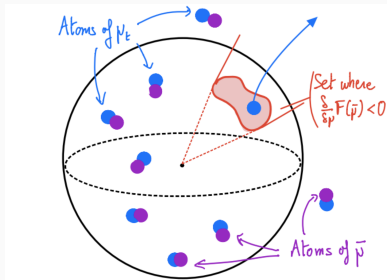
Theorem (Bach '14, reformulated)

For a suitable choice of regularization $\lambda(n) > 0$, the minimizer of F with square loss enjoys an error bound in $\tilde{O}(n^{-1/(r+3)})$.

- methods with fixed features (e.g. kernels) remain $\sim n^{-1/d}$
- no need to bound the number m of units

Proof Intuition (Global Convergence Thm.)

- Using homogeneity & convexity, μ^* minimizes F iff
 - (i) $\frac{\delta}{\delta\mu} F(\mu^*)[w] = 0$, for μ^* -a.e w (\Leftrightarrow stationary point of PDE)
 - (ii) $\frac{\delta}{\delta\mu} F(\mu^*)[w] \geq 0$
- If μ_0 has mass in all directions, so does μ_t for any $t \geq 0$
- If $\bar{\mu}$ does not satisfy (ii), then μ_t cannot be trapped near $\bar{\mu}$



- Thus if μ_t converges, its limit is a minimizer of F by (i) & (ii)