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

Lénaïc Chizat*, joint work with Francis Bach+

May 8th 2020 - MSRI Hot Topics

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

Supervised learning with neural networks

Supervised machine learning

- Consider a 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 such that $h(X) \approx Y$

Wide 2-layer ReLU neural networks

Class of predictors h of the form, for some large *width* $m \in \mathbb{N}$,

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

where $\phi(w, x) := c \max\{a^\top x + b, 0\}$ and $w := (a, b, c) \in \mathbb{R}^{d+2}$.

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

Learning algorithm: selects $(w_j)_j$ using the training data

Gradient flow of the empirical risk

Empirical risk minimization

- Choose a loss $\ell : \mathbb{R}^2 \to \mathbb{R}$ convex & smooth in its 1^{st} variable
- "Minimize" the empirical risk with a regularization $\lambda \geq 0$

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

Gradient-based learning

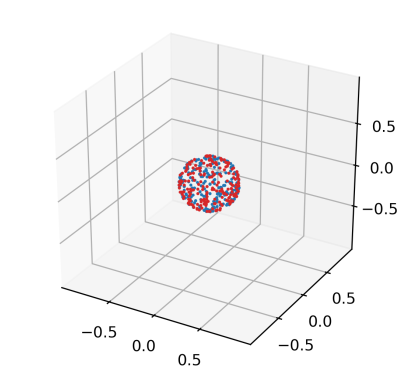
- Initialize $w_1(0), \dots, w_m(0) \overset{\text{i.i.d}}{\sim} \mu_0 \in \mathcal{P}_2(\mathbb{R}^{d+2})$
- Decrease the non-convex objective via gradient flow, for $t \geq 0$,

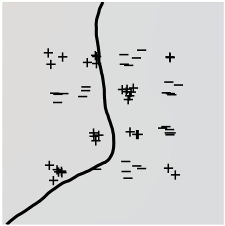
$$\frac{\mathrm{d}}{\mathrm{d}t}(w_j(t))_j = -m\nabla F_m((w_j(t))_j)$$

$\rightsquigarrow$ in practice, discretized with variants of gradient descent

Illustration

Dynamics for a classification task: unregularized logistic loss, $d = 2$





Space of parameters

- plot $|c| \cdot (a, b)$
- color depends on sign of c
- tanh radial scale

Space of predictors

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

Motivations

Main question

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

- Understanding 2-layer networks: when are they powerful?
  - $\rightsquigarrow$ role of initialization μ_0, loss, regularization, data structure, etc.
- Understanding representation learning via back-propagation
  - $\rightsquigarrow$ not captured by current theories for deeper models who study perturbative regimes around the initialization
- Natural next theoretical step after linear models
  - $\rightsquigarrow$ we can't understand the deep if we don't understand the shallow
- Beautiful connections with rich mathematical theories
  - $\rightsquigarrow$ variation norm spaces, Wasserstein gradient flows

Outline

Global convergence in the infinite width limit

Generalization with regularization

Implicit bias in the unregularized case

Wasserstein-Fisher-Rao gradient flows for optimization

Global convergence in the infinite width limit

Wasserstein gradient flow formulation

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

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

- Objective on the space of probability measures

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

Theorem (dynamical infinite width limit, adapted to ReLU)

Assume that

$$\mathrm{spt}(\mu_0) \subset \{|c|^2 = \|a\|_2^2 + |b|^2\}.$$

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

[Refs]:

Ambrosio, Gigli, Savaré (2008). *Gradient flows: in metric spaces and in the space of probability measures.*

Global convergence

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

Assume that $\mu_0 = \mathcal{U}_{\mathbb{S}^d} \otimes \mathcal{U}_{\{-1,1\}}$. If μ_t converges to μ_∞ in $\mathcal{P}_2(\mathbb{R}^{d+2})$, then μ_∞ is a global minimizer of F.

- Initialization matters: the key assumption on μ_0 is *diversity*
- Corollary: $\lim_{m,t\to\infty} F(\mu_{m,t}) = \min F$
- Convergence of μ_t: open question (even with compactness)

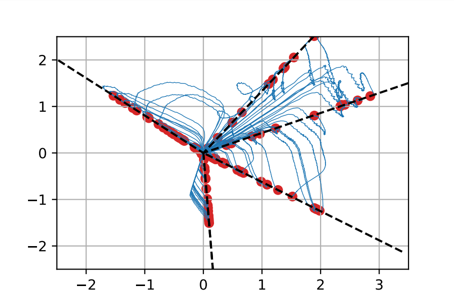
Generalization bounds?

They depend 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)

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

Illustration: population risk



Stochastic gradient descent on population risk ($m = 100$, $d = 1$)

Teacher-student setting: $X \sim \mathcal{U}_{\mathbb{S}^d}$ and $Y = f^*(X)$ where f^* is a ReLU neural network with 5 units (dashed lines)

Square loss $\ell(y, y') = (y - y')^2$.

[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.*

Generalization with regularization

Variation norm

Definition (Variation norm)

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

$$\|h\|_{\mathcal{F}_1} := \min_{\mu \in \mathcal{P}_2(\mathbb{R}^{d+2})} \left\{ \frac{1}{2} \int \|w\|_2^2 \, \mathrm{d}\mu(w) \; ; \; h(x) = \int \phi(w,x) \, \mathrm{d}\mu(w) \right\}$$

$$= \min_{\nu \in \mathcal{M}(\mathbb{S}^d)} \left\{ \|\nu\|_{TV} \; ; \; h(x) = \int \max\{a^\top x + b, 0\} \, \mathrm{d}\nu(a,b) \right\}$$

Proposition

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

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

[Refs]:

Neyshabur, Tomioka, Srebro (2015). *Norm-Based Capacity Control in Neural Networks.*

Kurkova, Sanguineti (2001). *Bounds on rates of variable-basis and neural-network approximation.*

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 $\mathbf{E}\big[(h(X) - f^*(X))^2\big]$ 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 $\ell(y, y') = (y - y')^2$ 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

[Refs]:
Bach. (2014). *Breaking the curse of dimensionality with convex neural networks.*

Fixing hidden layer and conjugate RKHS

What if we only train the output layer?

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

Definition (Conjugate RKHS)

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

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

Proposition (Kernel ridge regression)

All else unchanged, fixing the hidden layer leads to minimizing

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

- Solving: $\mathcal{F}_2$ random features, convex optim. / $\mathcal{F}_1$ difficult
- Priors: $\mathcal{F}_2$ isotropic smoothness / $\mathcal{F}_1$ anisotropic smoothness

Implicit bias in the unregularized case

Preliminary: linear classification and exponential loss

Classification task

- $Y \in \{-1, 1\}$ and the prediction is $\text{sign}(h(X))$
- $\ell(y, y') = \exp(-y'y)$ or logistic $\ell(y, y') = \log(1 + \exp(-y'y))$
- no regularization ($\lambda = 0$)

Theorem (Soudry et al. 2018, reformulated)

Consider $h(w, x) = w^\intercal x$ 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

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

[Refs]:
Soudry, Hoffer, Nacson, Gunasekar, Srebro (2018). *The Implicit Bias of Gradient Descent on Separable Data*.

Interpretation as online optimization

- look at $w'(t) = \nabla F_1(w(t))$, where F_β is the *smooth-margin*:

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

- prove that $\|w(t)\| \to \infty$ if the training set is linearly separable
- 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 on the sequence of objectives $F_{\|w(t)\|}$ which converge to the margin.

Implicit bias of two-layer neural networks

Let us go back to wide two-layer ReLU neural networks.

Theorem (C. & Bach, 2020)

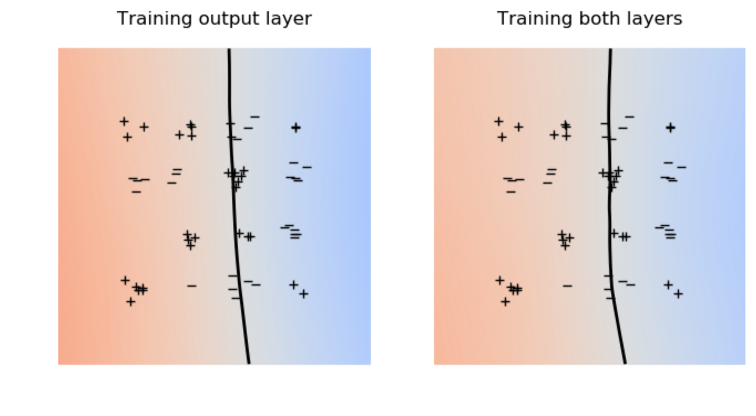
Assume that $\mu_0 = \mathcal{U}_{\mathbb{S}^d} \otimes \mathcal{U}_{\{-1,1\}}$, that the training set is consistant ($[x_i = x_j] \Rightarrow [y_i = y_j]$) and that μ_t and $\nabla F(\mu_t)$ converge in direction (i.e. after normalization). Then $h(\mu_t, \cdot)/\|h(\mu_t, \cdot)\|_{\mathcal{F}_1}$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

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

- no efficient algorithm is known to solve this problem
- fixing the hidden layer leads to the $\mathcal{F}_2$-max-margin classifier

[Refs]:
Chizat, Bach. *Implicit Bias of Gradient Descent for Wide Two-layer Neural Networks [...].*

Illustration



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

Statistical efficiency

Assume that $\|X\|_2 \leq R$ 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),

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

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

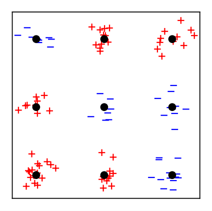
[Refs]:
Chizat, Bach. *Implicit Bias of Gradient Descent for Wide Two-layer Neural Networks [...].*

Numerical experiments

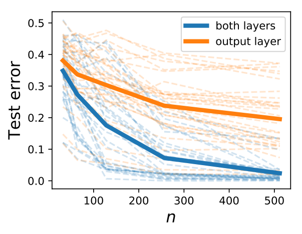
Setting

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

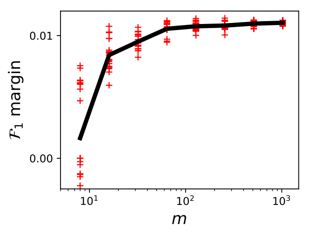
- two first coordinates as shown on the right
- all other coordinates uniformly at random



Coordinates 1 & 2



(a) Test error vs. n



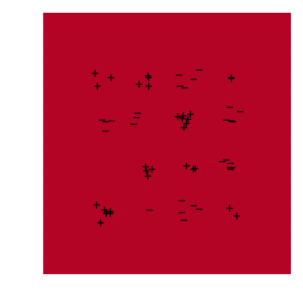
(b) Margin vs. m ($n = 256$)

Two implicit biases in one dynamics

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.

- Each neuron hardly moves but the total change in $h(\mu_t, \cdot)$ is significant
- 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



[Refs]:
Jacot, Gabriel, Hongler (2018). *Neural Tangent Kernel: Convergence and Generalization in Neural Networks.*
Chizat, Oyallon, Bach (2018). *On Lazy Training in Differentiable Programming.*
Woodworth et al. (2019). *Kernel and deep regimes in overparametrized models.*

Wasserstein-Fisher-Rao gradient flows for optimization

Convex optimization on measures

Definition (2-homogeneous projection)

Let $\Pi_2 : \mathcal{P}_2(\mathbb{R}^{d+2}) \to \mathcal{M}_+(\mathbb{S}^{d+1})$ satisfy $\forall\ \phi \in \mathcal{C}(\mathbb{R}^{d+2})$ 2-hmgn.:

$$\int_{\mathbb{R}^{d+2}} \phi(w)\,\mathrm{d}\mu(w) = \int_{\mathbb{S}^{d+1}} \phi(\theta)\,\mathrm{d}\Pi_2[\mu](\theta)$$

- With $\nu = \Pi_2[\mu]$, notice that we were in fact solving:

$$\min_{\nu\in\mathcal{M}_+(\mathbb{S}^{d+1})} J(\nu) = R\left(\int_{\mathbb{S}^{d+1}} \Phi(\theta)\,\mathrm{d}\nu(\theta)\right) + \lambda\nu(\mathbb{S}^{d+1})$$

- Inspiration to solve general convex optimization on measures?

Convex optimization on measures

Let Θ a d-manifold, J convex and with enough regularity. Solve

$$\min_{\nu\in\mathcal{M}_+(\Theta)} J(\nu).$$

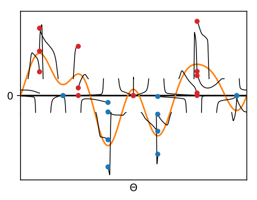
Conic Particle Gradient Descent

Algorithm (conic particle gradient descent)

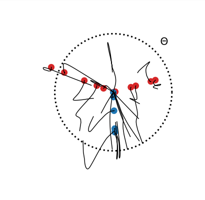
For $\alpha, \beta > 0$, discretize (with retractions) the gradient flow

$$\begin{cases} r_i'(t) = -4 r_i(t) J'_{\nu_t}(\theta_t(t)) \\ \theta_i'(t) = -\nabla J'_{\nu_t}(\theta_i(t)) \end{cases}$$

where $\nu_t = \frac{1}{m}\sum_{i=1}^m r_i(t)\delta_{\theta_i(t)}$ and J'_ν is the differential of J at ν.



(a) Sparse deconv.



(b) 2-layer neural net.

$\rightsquigarrow$ equivalent to gradient flow on a 2-layer ReLU neural net

Some properties

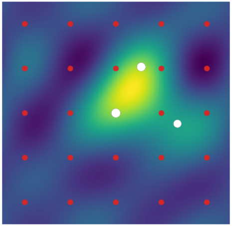
- ν_t is a Wasserstein-Fisher-Rao gradient flow of J, i.e. solves

$$\partial_t \nu_t = -\mathrm{div}\left(-\nabla J'_{\nu_t}\nu_t\right) - 4J'_{\nu_t}\nu_t$$

- If ν_0 has full support and $\nu_t \rightharpoonup \nu_\infty$ then ν_∞ minimizes J

For "non-degenerate sparse" problems:

- local exponential convergence
- ϵ-accurate solution in $O(\log(1/\epsilon))$
- but m exponential in d so far



Sparse deconvolution (white) sources (red) particles.

[Refs]

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

Conclusion

- **Generalization guarantees for gradient methods on neural nets**
- **Analysis via Wasserstein gradient flow with homogeneity**

Perspectives

- **Proof of convergence, quantitative results**
- **More complex architectures**

[Papers :]
- Chizat and Bach (2018). On the Global Convergence of Over-parameterized Models using Optimal Transport
- Chizat (2019). Sparse Optimization on Measures with Over-parameterized Gradient Descent
- Chizat, Bach (2020). Implicit Bias of Gradient Descent for Wide Two-layer Neural Networks Trained with the Logistic Loss