Scidolyse
Groupe de travail
Mathematics of Deep Learning

On the convergence of gradient descent for 2-layer neural networks: A mean-field view

Src: Andrea Montanari (course, '18, '19)
Chizat and Bach '19 (Neurips)

3 challenges:

- Approximation
- Convergence of Gradient Descent
- Generalization
  (no over-fitting)

Supervised Learning

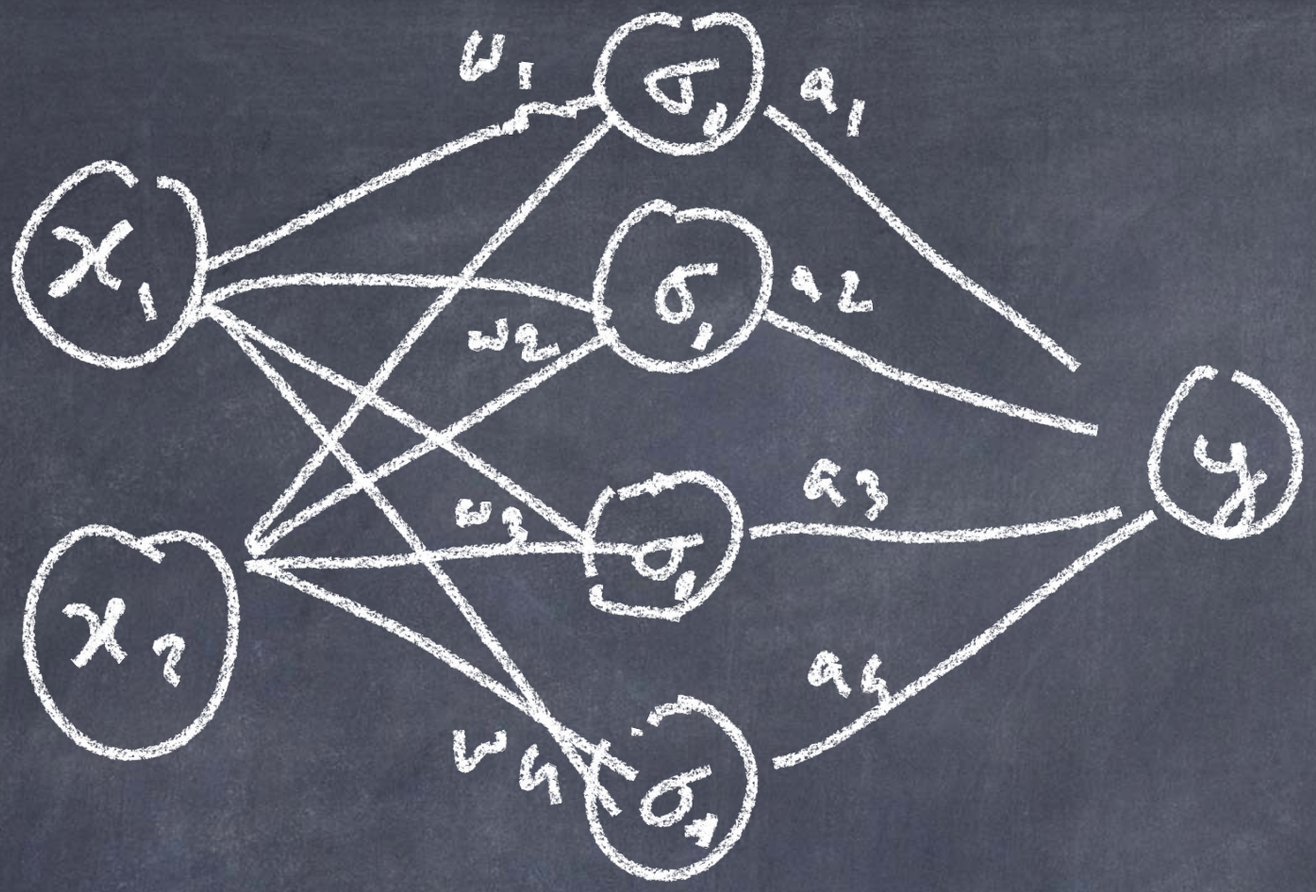
- $(X, Y) \sim P$
- sample $(X_1, Y_1) \dots (X_n, Y_n)$
- $R(\theta) = \mathbb{E}\left[(f(X;\theta) - Y)^2\right]$

$$\hat{R}(\theta) = \mathbb{E}_n\left[(f(X;\theta) - Y)^2\right]$$

$$= \frac{1}{n} \sum_{i=1}^{n} (f(x_i, \theta) - y_i)^2$$

Neural Nets

1 hidden layer



$$y = f(x;\theta) := \frac{1}{N}\sum_{i=1}^{N} \sigma(x;\theta_i)$$

$$\theta_i = (a_i, b_i, \omega_i)$$

$$\sigma(x,\theta) = a\,\sigma_0(\langle \omega, x\rangle + b)$$

$$\sigma_0(x) = x_+ \quad \text{or} \quad \sigma_0(x) := \frac{1}{1+e^{-2x}} \quad \dots$$

Th [Cybenko '89]

If $\int \|f(x)\|^2 < \infty$

If $\sigma : \mathbb{R} \to \mathbb{R}$ is continuous, $\begin{cases} \sigma(x) \xrightarrow[x \to +\infty]{} 1 \\ \sigma(x) \xrightarrow[x \to -\infty]{} 0 \end{cases}$

Then $\forall \varepsilon > 0$, $\exists N(\varepsilon)$ s.t.

$$\inf_{(a_i, b_i, \omega_i)_{1 \le i \le N}} \int \left[f(x) - \frac{1}{N} \sum_{i=1}^{N} a_i \sigma(\langle \omega_i, x \rangle + b_i) \right] \le \varepsilon$$

PC: how to find the good approximator?

→ Gradient Descent: (stepsize)

$$\theta^{k+1} = \theta^k + s_k v_k$$

$$v_k = -\nabla R(\theta^k)$$

SGD:

$$v_k = -\nabla R(\theta^k) + \underbrace{\varepsilon^k}_{\text{noise}}$$

TB: $\tilde{R}(\theta)$ is not convex in θ
it has local minima, etc...

BUT
still (S)GD works
especially when the network is
over-parameterized -> WHY?

Here: $R_N(\theta) = \mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N} \sigma(x;\theta) - y\right)^2\right]$

$\rightarrow \theta_i^{k+1} = \theta_i^k + 2\eta \mathbb{E}\left[\sigma(x,\theta_i)\left(y - \frac{1}{N}\sum_{j=1}^{N} \sigma(x,\theta_j)\right)\right]$

$R_N(\theta) = \mathbb{E}[y^2] + \frac{2}{N}\sum_{i=1}^{N} -\mathbb{E}[y\sigma(x,\theta_i)] + \frac{1}{N^2}\sum_{i,j=1}^{N} \mathbb{E}[\sigma(x,\theta_i)\sigma(x,\theta_j)]$

$R(\theta) = R_\# + \frac{2}{N}\sum_{i=1}^{N} V(\theta_i) + \frac{1}{N^2}\sum_{i,j=1}^{N} U(\theta_i,\theta_j)$

$R(\theta)$: energy

$V(\theta_i)$: external potential

$U(\theta_i,\theta_j)$: pairwise potential

Important: the kernel U is semi-definite:

$\forall h$ bounded, compactly supported

$$\iint U(\theta_1, \theta_2)\, h(\theta_1)\, h(\theta_2)\, d\theta_1\, d\theta_2 \geq 0$$

repulsive interaction (in average sense)

Convexification

$$f(x,\theta) = \frac{1}{N}\sum_{i=1}^{N} \sigma(x,\theta_i)$$

$$= \int \sigma(x,\theta)\, \rho(d\theta)$$

when N is large, can be modelled by a density ρ

$\to R_N(\theta)$

$\geq R(\rho) : R_{\#} + 2\int V(\theta)\rho(d\theta)$

$+ \int U(\theta_1,\theta_2)\rho(d\theta_1)\rho(d\theta_2)$

Prop: if $\exists \varepsilon_0 : \forall \rho,\ R(\rho) \leq \inf R(\rho) + \varepsilon_0$

$\Rightarrow \int U(\theta,\theta)\rho(d\theta) \leq K$

then $\left|\inf_\theta R_N(\theta) - \inf_\rho R(\rho)\right| \leq \frac{K}{N}$

Remark:

$$R(\rho) = R_\# + 2\int V(\theta)\rho(d\theta) + \int V(\theta_1,\theta_2)\rho(d\theta_1)\rho(d\theta_2)$$

is convex in ρ !

$\Rightarrow$ convex optimization in ∞-dim coll. (Θ)

Δ- Sum approach

Functional derivative

$$\psi(\theta, \rho) = \frac{1}{2} \frac{\delta R(\rho)}{\delta \rho(\theta)} = V(\theta) + \int U(\theta, \theta') \rho(d\theta')$$

= variation of energy when adding 1 particle at θ

ρ_* is a minimum if

$$\operatorname{supp}(\rho_*) \subset \underset{\theta \in \mathbb{R}^3}{\operatorname{argmin}} \ \psi(\theta, \rho_*)$$

→ Idea 1:

- Discretize $\Omega \to \Omega_N$
- Minimize $R(\rho_N)$ on $M_1(\Omega_N) \subseteq \mathbb{R}^N$

$$R(\rho_N) = \langle R_{\#}\, \rho_N \rangle + \rho_N V_N + \rho_N U_N \rho_N'$$

where $V_N(i) := V(\theta_i)$ and $U_N(i,j) := u(\theta_i, \theta_j)$

Pb: curse of dimensionality:

requires gigantic N.

→ What follows proves that a good N does not need to be exponential in D.

Idea 2: particular approach

$$\rho_N = \frac{1}{N} \sum_{i=1}^{N} \delta_{\theta_i}$$

each particle θ_i moves according to the force of the system: at time t the speed $\pi_t(\dot{\theta}_t^i \mid \mathcal{F}_t) = -\nabla_{\theta_i} V(\theta_t^i) - \frac{1}{N} \sum_{j=1}^{N} \nabla_{\theta_i} u(\theta_t^i, \theta_t^j)$

$$= \mathbb{E}[\nabla \sigma(x, \theta_t^i)] - \frac{1}{N} \sum_{j=1}^{N} \nabla_{\theta_i} \mathbb{E}[\sigma(x, \theta_t^i) \sigma(x, \theta_t^j)]$$

$\rightarrow$ this is exactly (S)GD

→ Prove that the particle system behaves like its continuous equivalent (statistical mechanics)

$$\mathbb{E}\left[\sigma_i^k \mid \mathcal{F}_k\right] = -\nabla \psi\left(\theta_i^k, \ell_{best}\right)$$

$$\left((\rho_t)_{t \geq 0}\right. \text{ continuous time limit.} \qquad t = k\varepsilon.$$

Continuity equation

$$\partial_t \rho_t = -\underbrace{\nabla \cdot}_{\text{div}} \left(\rho_t(\theta)\, v(\theta, \rho_t) \right)$$

enters: $\rho_t(\theta)\, v(\theta, \rho_t)$

leaves: $\rho_t(\theta + d\theta)\, v(\theta + d\theta, \rho_t)$

$\Rightarrow$ variation of ρ_t: $\frac{\partial}{\partial \theta} \left(\rho_t(\theta)\, v(\theta, \rho_t) \right)$

$$\rightarrow \partial_t \rho_t(\theta) = \nabla_\theta \cdot \left(\rho_t(\theta) \nabla_\theta \psi(\theta, \rho_t) \right)$$

Fixed points = densities ρ_* s.t.
all mass sits on zero velocity points:

$$\text{supp}(\rho_*) \subset \{\theta \in \mathbb{R}^D : \nabla \psi(\theta, \rho_*) = 0\}$$

Thm: if $\|\sigma_*\|_\infty \le K_2$, $\|\nabla_x \sigma_*(x;\theta)\|_2 \le K_2$
$|y_u| \le K_2$
if $\|\nabla_x v(\theta)\|_2 \le K_3$, $\|\nabla_{\theta_1,\theta_2} v(\theta_1,\theta_2)\|_2 \le K_3$
$\|\nabla_\theta v(\theta) - \nabla_\theta v(\theta')\|_2 \le K_3 \|\theta - \theta'\|_2$
$\|\nabla_{\theta_1,\theta_2} v(\theta_1,\theta_2) - \nabla_{\theta_1,\theta_2} v(\theta_1,\theta_2')\|_2 \le K_3 \left\|\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix} - \begin{pmatrix}\theta_1'\\ \theta_2'\end{pmatrix}\right\|_2$
$\rho_0 \in \mathcal{P}(\mathbb{R}^D)$. SGD with initialization $(\theta_i^0)_{1\le i\le N} \sim \rho_0$
step size $\gamma_k = \frac{\varepsilon}{2}$
ρ_t : solution of the TDE
Then $\exists C = C(K_i)$, $\forall f : \mathbb{R}^P \times \mathbb{R} \to \mathbb{R}$ s.t. $\|f\|_{\mathrm{Lip}} \le 1$, $\|f\|_\infty \le 1$
$\sup_{k=0}^{T/\varepsilon} \left|R_N(\theta^k) - R(\rho_{k\varepsilon})\right| \le C e^{CT} \sqrt{\frac{1}{N}\vee\varepsilon}\left[\sqrt{D + \log\frac{T}{\varepsilon}} + 3\right]$
with proba $\ge 1 - e^{-3^2}$

→ TDF accurate as $N \gg D$, $\varepsilon \ll 1/D$

→ no curse of dimensionality.

numerical experiments → PDE approx very accurate in practice.

→ global convergence can be proved in some cases.

Gradient Flows

$$\dot{x}(t) = -\nabla F(x(t))$$

"continuous time gradient descent"

$$x(t+\varepsilon) = \operatorname*{argmin}_{z} \left\{ F(z) + \frac{1}{2\varepsilon} \|z - x(t)\|_2^2 \right\}$$

cf GD: $x_{k+1} = \operatorname*{argmin}_{z} \left\{ F(z) + \frac{1}{2\eta_k} \|z - x_k\|^2 \right\}$

→ general definition: for a distance $d(\cdot,\cdot)$

$$x_\varepsilon((k+1)\varepsilon) := \operatorname*{argmin}_{\xi \in \pi_X} \left\{ F(\xi) + \frac{1}{2\varepsilon} d(\xi, x_\varepsilon(t))^2 \right\}$$

$$x(t) := \lim_{\varepsilon \to 0} x_\varepsilon(t)$$

is the gradient flow of the cost function F on X for the metric d

Prop:

$$\partial_t \rho(t) = \nabla_\theta \cdot (\rho_r(\theta) \nabla_\theta \psi(\theta, \rho_t))$$

is the gradient flow for the cost $R(\theta)$ in Wasserstein metric

$$W_2(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int \|x - y\|_2^2 \, \gamma(dx, dy) \right)^{1/2}$$

$\Gamma(\mu,\nu)$: couplings of μ and ν: probs γ on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals μ and ν

Noisy GD

$$\theta_i^{k+1} = \theta_i^k + 2\varepsilon_k \nabla_{\theta_i} \sigma(x_k, \theta_i^k)\left(y_k - \frac{1}{N}\sum_{i=1}^{N} \sigma(x_k, \theta_i^k)\right) + \sqrt{2\varepsilon_k}\, g_i^k$$

$$g_i^k \sim \mathcal{N}(0, I_D)$$

$$\rightarrow \partial_t \rho_t(\theta) = \nabla_\theta \cdot \left(\rho_t(\theta) \nabla_\theta \Psi(\theta, \rho_t)\right) + \tau \Delta \rho_t(\theta)$$

$$F(\rho) = \frac{1}{2} R(\rho) - \tau S(\rho) \quad \text{free energy}$$

where $S(\rho) = -\int \rho(\theta) \log \rho(\theta)\, d\theta$ entropy

$$\rho_*(\theta) = \frac{1}{Z(\beta)} \exp(-\beta \Psi(\theta, \rho_*)) \quad \text{Boltzmann}$$

One can prove convergence in a time that depends on D but not on N -

$\rightarrow$ SGD reaches a near-optimum in time independent of the number of neurons