Gradient descent dynamics. Part 2

Theoretical Deep Learning course

Eugene Golikov

MIPT, spring 2019

Neural Networks and Deep Learning Lab., MIPT

Local optimization: neural nets?

Are results all of these results applicable to neural nets?

- For ReLU nets $\mathcal{L}(W) \notin \mathcal{C}^2$;
- $\mathcal{L}(W)$ has non-strict saddles even for linear networks of depth ≥ 3;
- We cannot guarantee convergence to a global minimum in reasonable amount of steps;
- SGD is neither NoisyGD, nor PGD;
- Convergence in $\log^4 d$ is still too slow.

Desired result for neural nets:
Given number of weights is large enough and learning rate is small enough, GD (or SGD) quickly converges to global minimum of $\mathcal{L}(W)$ *with high probability over random initialization.*

GD on non-linear nets

A ReLU-net (Du et al., 2018a[1]):

$$f(W, a, x) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r [w_r^T x]_+, \quad x \in \mathbb{R}^d, \, w_r \in \mathbb{R}^d, \, a_r \in \mathbb{R}.$$

$$L(W) = \frac{1}{2} \sum_{i=1}^{n} (f(W, a, x_i) - y_i)^2.$$

Continuous-time GD:

$$\frac{dw_r(t)}{dt} = -\frac{\partial L(W(t))}{\partial w_r}.$$

[1] https://openreview.net/forum?id=S1eK3i09YQ

GD on non-linear nets

Some assumptions and definitions:

- Initialization:

$$w_r \sim \mathcal{N}(0, I),\ a_r \sim U(\{-1, 1\}) \quad \forall r = 1, \dots, m;$$

- Data points: $\|x_i\|_2 = 1,\ |y_i| < C \quad \forall i = 1, \dots, n;$
- Individual predictions: $u_i(t) := f(W(t), a, x_i);$
- Gram matrix:

$$H_{ij}(W) := \frac{1}{m} \sum_{r=1}^{m} (x_i^T x_j [w_r^T x_i \geq 0, w_r^T x_j \geq 0]) \quad \forall i, j = 1, \dots, n;$$

- Mean Gram matrix at initialization:

$$H_{ij}^{\infty} := \mathbb{E}_{w \sim \mathcal{N}(0,I)} (x_i^T x_j [w^T x_i \geq 0, w^T x_j \geq 0]) \quad \forall i, j = 1, \dots, n;$$

- $\lambda_0 := \lambda_{min}(H^{\infty}).$

GD on non-linear nets

Proof road-map:

- **Lemma 3.1:** *for large enough* m $\lambda_{min}(H(0)) \geq \frac{3}{4}\lambda_0$ *w.h.p.*
- **Lemma 3.2:** *for all* W *sufficiently close to* $W(0)$ $\lambda_{min}(H) \geq \frac{1}{2}\lambda_0$ *w.h.p.*
- **Lemma 3.3:** *if* $\lambda_{min}(H(s)) \geq \frac{1}{2}\lambda_0$ $\forall s \in [0, t]$, *then* $W(t)$ *is sufficiently close to* $W(0)$.
- **Lemma 3.4:** *for large enough* m *for all* $t \geq 0$ $\lambda_{min}(H(t)) \geq \frac{1}{2}\lambda_0$ *and* $W(t)$ *is sufficiently close to* $W(0)$ *w.h.p.*

GD on non-linear nets

Theorem (Du et al., 2018a):

Let $\delta \in (0,1)$ and $m = \Omega\left(\frac{n^6}{\lambda_0^4 \delta^3}\right)$; then w.p. $\geq 1-\delta$ over initialization we have

$$\|u(t) - y\|_2^2 \leq e^{-\lambda_0 t}\|u(0) - y\|_2^2 \quad \forall t \geq 0.$$

Extensions:

- **Training both layers:** bound on m weakens: $m = \Omega(\frac{n^6 \log(m/\delta)}{\lambda_0^4 \delta^3})$.
- **Discrete-time GD:** For $\delta \in (0,1)$, $m = \Omega\left(\frac{n^6}{\lambda_0^4 \delta^3}\right)$, and step size $\eta = O(\lambda_0/n^2)$, w.p. $\geq 1-\delta$ over initialization we have:

$$\|u(k) - y\|_2^2 \leq \left(1 - \frac{\eta\lambda_0}{2}\right)^k \|u(0) - y\|_2^2 \quad \forall k \geq 0.$$

A non-linear net with one hidden layer:

$$\mathcal{L}(W) = \|Y - W_2\sigma(W_1X)\|_F^2,$$

where $X \in \mathbb{R}^{d_0 \times m}$, $W_1 \in \mathbb{R}^{d_1 \times d_0}$, $W_2 \in \mathbb{R}^{d_2 \times d_1}$ and $Y \in \mathbb{R}^{d_2 \times m}$.

Theorem Yu & Chen (1995)[2]:

Suppose

1. $\sigma(z) = (1 + \exp(-z))^{-1}$,
2. all columns of X are distinct,
3. $d_1 = m$.

Then all local minima of $\mathcal{L}$ are global.

[2] https://ieeexplore.ieee.org/document/410380

GD on non-linear nets

A deep net with smooth activations (Du et al., 2018b[3]):

$$f(W, a, x) = a^T \sqrt{\frac{c_\sigma}{m}} \sigma \left(W^{(H)} \sqrt{\frac{c_\sigma}{m}} \sigma \left(W^{(H-1)} \dots \sqrt{\frac{c_\sigma}{m}} \sigma \left(W^{(1)} x \right) \right) \right),$$

where c_σ is a constant determined by the activation σ.

$$\mathcal{L}(W) = \frac{1}{2} \sum_{i=1}^{n} (f(W, a, x_i) - y_i)^2.$$

Initialization strategy and data assumptions are the same as in Du et al. (2018a).

[3] `https://arxiv.org/abs/1811.03804`

GD on non-linear nets

Theorem (Du et al., 2018b):

Let $\delta \in (0,1)$, $m = \Omega\left(\max\left(\frac{n^4 2^{O(H)}}{\lambda_0^4}, \frac{n2^{O(H)}}{\delta}, \frac{n^2 \log(Hn^2/\delta)}{\lambda_0^2 \lambda^{3H/2}}\right)\right)$ and $\eta = O(\frac{\lambda_0}{n^2 2^{O(H)}})$; then wp $\geq 1-\delta$ over initialization, for k-th step of GD we have

$$\mathcal{L}(W_k) \leq \left(1 - \frac{\eta\lambda_0}{2}\right)^k \mathcal{L}(W_0).$$

Remarks:

- Essentially the same proof strategy as in Du et al. (2018a);
- (Almost surely) could be generalized to ReLU nets (with larger bound on m);
- Exponential dependence on the number of layers H.