Gradient descent dynamics

Theoretical Deep Learning course

Eugene Golikov

MIPT, spring 2019

Neural Networks and Deep Learning Lab., MIPT

Brief overview

Objective:

$$\mathcal{L}(W) = \mathbb{E}_{x,y\sim\mathcal{D}} L(y, \hat{y}(x, W)) \to \min_W,$$

where W – network weights, $\hat{y}$ – network response, $\mathcal{D}$ – true data distribution, L – loss function.
Dimension of $W > 10^4$ (typically $10^6 \div 10^8$).

- **Previous lecture:** Study critical points of $\mathcal{L}(W)$
- **This lecture:** Study dynamics of gradient descent on $\mathcal{L}(W)$

Brief overview

Previous lecture:

In some cases (linear nets, wide non-linear nets) all local minima are guaranteed to be global.

Questions:

1. Why don't we converge to saddle points?
2. How fast do we converge to minima?
3. Even if there are non-global minima, how often do we converge to them?

Plan

1. General convergence guarantees:
    1.1 Gradient descent
    1.2 Gradient descent with random init
    1.3 Noisy gradient descent
2. Linear nets
3. Non-linear nets:
    3.1 Shallow ReLU-nets
    3.2 Extensions to deep nets

Local optimization: general case

Let $f : \mathbb{R}^d \to \mathbb{R}$.

Definition:

We write $f \in \mathcal{C}_L^{k,q}$ for $q \leq k$, if:

- $f \in \mathcal{C}^k(\mathbb{R}^d)$;
- $\|\nabla^q f(x) - \nabla^q f(y)\|_2 \leq L\|x - y\|_2 \quad \forall x, y$.

Stationary points:

- For $f \in \mathcal{C}^1$ x^* is a 1st-order stationary point if:

$$\|\nabla f(x^*)\|_2 = 0;$$

- For $f \in \mathcal{C}_\rho^{2,2}$ x^* is a 2nd-order stationary point if:

$$\|\nabla f(x^*)\|_2 = 0, \qquad \lambda_{min}(\nabla^2 f(x^*)) \geq 0.$$

Local optimization: general case

Let $f:\ \mathbb{R}^d \to \mathbb{R}$.

Definition:

We write $f \in \mathcal{C}_L^{k,q}$ for $q \leq k$, if:

- $f \in \mathcal{C}^k(\mathbb{R}^d)$;
- $\|\nabla^q f(x) - \nabla^q f(y)\|_2 \leq L\|x-y\|_2 \quad \forall x, y.$

Stationary points:

- For $f \in \mathcal{C}^1$ x^* is an ϵ-1st-order stationary point if:

$$\|\nabla f(x^*)\|_2 \leq \epsilon;$$

- For $f \in \mathcal{C}_\rho^{2,2}$ x^* is an ϵ-2nd-order stationary point if:

$$\|\nabla f(x^*)\|_2 \leq \epsilon, \qquad \lambda_{min}(\nabla^2 f(x^*)) \geq -\sqrt{\rho\epsilon}.$$

Local optimization: general case

Suppose $f \in \mathcal{C}^1$, $f \geq 0$.

Gradient descent:

$$x_{t+1} = x_t - \alpha \nabla f(x_t).$$

Guarantees:

Given $f \in \mathcal{C}_L^{1,1}$, $\alpha \in (0, 2/L)$ and any x_0:

- $f(x_t) \searrow$ (hence $f(x_t) \to f^*$);
- GD achieves an ϵ-1st-order stationary point in

$$T_\epsilon = \epsilon^{-2} \frac{L}{\omega(\alpha)} (f(x_0) - f^*) \quad \text{iterations.}$$

No guarantees to converge to ϵ-2nd-order stationary point in non-convex case!

Local optimization: general case

Ways to guarantee convergence to 2nd-order stationary point:

1. Introduce random initialization (Lee et al., 2016[1]):

$$x_0 \sim P_{init}(x_0);$$

2. Introduce noise to gradients (Ge et al., 2015[2], Jin et al., 2017[3]):

$$x_{t+1} = x_t - \alpha(\nabla f(x_t) + \xi_t), \quad \xi_t \sim P_{noise}(\xi).$$

[1] https://arxiv.org/abs/1602.04915
[2] https://arxiv.org/abs/1503.02101
[3] https://arxiv.org/abs/1703.00887

Local optimization: general case

Suppose $f \in \mathcal{C}_L^{2,1}$, $f \geq 0$.

Theorem (Lee et al., 2016):

Let $x_0 \sim P_{init}(x_0)$. Assume P_{init} is absolutely continuous wrt Lebesgue measure μ.
Then for any critical point x^* only one of two holds:

- x^* is a 2nd-order stationary point;
- $\mathbb{P}(\lim x_t = x^*) = 0$.

Equivalently, if x^* is a *strict saddle*, i.e. $\lambda_{min}(\nabla^2 f(x^*)) < 0$, then $\mathbb{P}(\lim x_t = x^*) = 0$.

Local optimization: general case

Global stable set:

We call $W^s(x^*)$ a *global stable set* of a critical point x^* if

$$W^s(x^*) = \{x_0 : \lim_{t \to \infty} x_t = x^*\}.$$

Trivial fact:

If $P_{init}(W^s(x^*)) = 0$ then GD almost surely doesn't converge to x^*.

Local optimization: general case

Rewrite GD dynamics as:

$$x_{t+1} = x_t - \alpha \nabla f(x_t) = g(x_t) = g^{t+1}(x_0).$$

Local stable set:

We call $W^s_{loc}(x^*)$ a *local stable set* of a critical point x^* if

$$\exists U \text{ – vicinity of } x^*\text{: } W^s_{loc}(x^*) = U \cap W^s(x^*).$$

We can reconstruct global stable set from the local one:

$$W^s(x^*) = \bigcup_{t=0}^{\infty} g^{-t}(W^s_{loc}(x^*)).$$

Local optimization: general case

$$W^s(x^*) = \bigcup_{t=0}^{\infty} g^{-t}(W^s_{loc}(x^*)).$$

P_{init} is absolutely continuous wrt Lebesgue measure μ on $\mathbb{R}^d$:

$$\mu(W) = 0 \Rightarrow P_{init}(W) = 0 \quad \forall W.$$

We want to prove that $\mu(W^s(x^*)) = 0$ for any strict saddle x^*.
For this, it is sufficient to have:

1. $\mu(W^s_{loc}(x^*)) = 0$;
2. g is a diffeomorphism.

Local optimization: general case

Proposition (Lee et al., 2016):

For $f \in \mathcal{C}_L^{2,1}$ the gradient mapping g with step size $\alpha \in (0, 1/L)$ is a diffeomorphism.

Theorem[4][5]:

Informally: for every stable point x^* of a diffemorphism g there exist an embedded disk W^s_{loc}, which is a local stable set of x^*;
$\dim W^s_{loc} = \#$ eigenvalues of $Jg(x^*)$ less or equal to 1 in absolute value.

[4] https://en.wikipedia.org/wiki/Stable_manifold_theorem
[5] https://en.wikipedia.org/wiki/Center_manifold

Local optimization: general case

Corollary (Lee et al., 2016):

Let x^* be a strict saddle, i.e. $\lambda_{min}(\nabla^2 f(x^*)) < 0$. Then,

$$\begin{aligned} \operatorname{codim} W^s_{loc} &= \#\text{ eigenvalues of } Jg(x^*) \text{ greater than 1 in abs. value} = \\ &= \#\text{ eigenvalues of } \nabla^2 f(x^*) \text{ less than } 0 > 0. \end{aligned}$$

Hence $\mu(W^s_{loc}(x^*)) = 0$. Consequently, $\mu(W^s(x^*)) = 0$ and GD almost surely doesn't converge to x^*.

Local optimization: general case

Theorem (Lee et al., 2016):

Let $x_0 \sim P_{init}(x_0)$. Assume P_{init} is absolutely continuous wrt Lebesgue measure μ.
Then, for any strict saddle x^*, $\mathbb{P}(\lim x_t = x^*) = 0$.

Question:

Suppose GD is guaranteed to converge to a 1st-order stationary point:

$$\forall x_0 \, \exists x^* : \quad \lim x_t = x^*, \; \nabla f(x^*) = 0.$$

Let all saddles of f be strict, and there are no maxima (*strict saddle property*).
Does the theorem guarantee that GD converges to a local minimum?

Local optimization: general case

Convergence time guarantees for randomly initialized GD?

1st-order stationary point:

GD finds an ϵ-1st-order stationary point of $f \in \mathcal{C}_L^{1,1}$ in time T_ϵ, independent from d:

$$T_\epsilon \propto \epsilon^{-2} L$$

2nd-order stationary point:

Generally, good guarantee is impossible:
Du et al. (2017)[6] constructed $f \in \mathcal{C}_L^{2,1}$ with strict saddle property, for which GD with random initialization almost surely finds an ϵ-2nd-order stationary point in time exponential wrt d.

[6] `https://arxiv.org/abs/1705.10412`

Local optimization: general case

Noisy gradient descent (NoisyGD):

$$x_{t+1} = x_t - \alpha(\nabla f(x_t) + \xi_t), \quad \xi_t \sim \mathrm{Unif}(\mathcal{S}^{d-1}(1)).$$

Theorem (Ge et al., 2015):

Suppose $f \in \mathcal{C}_L^{2,1} \cap \mathcal{C}_\rho^{2,2}$, $|f(x)| \leq B \; \forall x \in \mathbb{R}^d$.
Then $\forall \; \epsilon > 0, \delta > 0 \; \exists \alpha_{max}(\epsilon, \delta) : \; \forall \alpha < \alpha_{max}$ NoisyGD achieves an ϵ-2nd-order stationary point in

$$T_\epsilon = O(poly(d/\epsilon)) \quad \text{iterations w.p. } 1 - \delta.$$

Remark:

What is necessary is sufficient variance in every direction $\Rightarrow$ can substitute gradient with stochastic gradient.

Local optimization: general case

Perturbed gradient descent (PGD):

Very informally:

- If gradient is large, do a GD step;
- If gradient is small, inject noise, then do several GD steps;
- Stop, if f didn't decrease sufficiently after some GD steps after injecting noise.

Theorem (Jin et al., 2017):

Suppose $f \in \mathcal{C}_L^{2,1} \cap \mathcal{C}_\rho^{2,2}$, $|f(x)| \leq B \ \forall x \in \mathbb{R}^d$.
Then $\forall \ \epsilon > 0, \delta > 0$ for appropriate choice of hyperparameters PGD achieves an ϵ-2nd-order stationary point in

$$T_\epsilon = O(log^4(d)/\epsilon^2) \quad \text{iterations w.p. } 1 - \delta.$$

Local optimization: general case

Let $f(x) = \frac{1}{m}\sum_{i=1}^m f_i(x)$.

Stochastic gradient descent (SGD):

$$x_{t+1} = x_t - \alpha \nabla f_i(x_t), \quad i \sim \mathrm{Unif}(\{1..m\}).$$

Since $\mathbb{E}_i \nabla f_i(x_t) = \nabla f(x_t)$:

$$\nabla f_i(x_t) = \nabla f(x_t) + \xi_t, \quad \mathbb{E}\xi_t = 0.$$

Stochastic GD $\approx$ Noisy GD?

If we want to apply theorem of Ge et al. (2015), we need to ensure ξ_t has variance bounded from below in every direction.
That's not true: see Chaudhari & Soatto (2018)[7].

[7] `https://openreview.net/forum?id=HyWrIgW0W&noteId=HyWrIgW0W`

Local optimization: general case

Stochastic gradient Langevin dynamics (SGLD):

$$x_{t+1} = x_t - \alpha_t(\nabla f(x_t) + \xi_t) + \beta_t W, \quad W \sim \mathcal{N}(0, 1),$$

where ξ_t is stochastic gradient noise, for which $\mathbb{E}\xi_t = 0$, $\mathbb{E}\xi_t^2 < Q$.

Theorem (Gelfand & Mitter, 1990[8]):

$\exists C_0$: for $\alpha_t = A/t$, $\beta_t = \sqrt{B/t \log\log t}$, where $B/A > C_0$, and *tight initialization strategy* P_{init},

$$x_t \overset{prob}{\to} x^*, \quad x^* \sim \pi,$$

where π is a weak limit of π^ϵ as $\epsilon \to 0$:

$$d\pi^\epsilon(x) = \frac{1}{Z^\epsilon} \exp\left(\frac{-2f(x)}{\epsilon^2}\right) dx.$$

[8] https://core.ac.uk/download/pdf/4380833.pdf

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.