Loss function landscape. Part 1

Theories of Deep Learning

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 $10^6 \div 10^8$).

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 $10^6 \div 10^8$).

Two questions arise:

1. **This topic:** How does loss function $\mathcal{L}(W)$ look like?
2. **Next topic:** Why does gradient descent perform well on this task?

Plan

1. General case:
    1.1 Impossibility of global optimization
    1.2 Local optimization
2. Simple cases:
    2.1 Deep linear nets
    2.2 Non-linear nets
3. Spherical spin-glass model
4. Back to general case

Neural network learning as black-box optimization

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 $10^6 \div 10^8$).

Consider $\mathcal{L}(W)$ is a black-box:

- Can compute local quantities ($\mathcal{L}(W)$, $\nabla\mathcal{L}(W)$ at any point W), but know nothing about global landscape.

[1] `https://link.springer.com/article/10.1007/BF02592948`

Neural network learning as black-box optimization

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 $10^6 \div 10^8$).

Consider $\mathcal{L}(W)$ is a black-box:

- Can compute local quantities ($\mathcal{L}(W)$, $\nabla\mathcal{L}(W)$ at any point W), but know nothing about global landscape.

Hence we have to explore the whole parameter space.

Global black-box optimization is known to be NP-complete!
See Murty & Kabadi (1987)[1]

[1] https://link.springer.com/article/10.1007/BF02592948

Neural network learning as black-box optimization

Objective:

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

Can we exploit the structure of NN?

$$\hat{y}(x, W) = W_H \sigma(W_{H-1} \ldots \sigma(W_1 x) \ldots).$$

[2] https://papers.nips.cc/paper/125-training-a-3-node-neural-network-is-np-complete.pdf

[3] www2.cs.cas.cz/~sima/trhard.ps

Neural network learning as black-box optimization

Objective:

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

Can we exploit the structure of NN?

$$\hat{y}(x, W) = W_H \sigma(W_{H-1} \dots \sigma(W_1 x) \dots).$$

Generally, no: it is still NP-complete even for very simple NN

See Blum & Rivest (1992)[2]; see also Sima (2002)[3].

---

[2] `https://papers.nips.cc/paper/125-training-a-3-node-neural-network-is-np-complete.pdf`

[3] `www2.cs.cas.cz/~sima/trhard.ps`

Local optimization

Critical point condition:

$$\nabla \mathcal{L}(W) = 0.$$

Critical points are characterized by the eigenvalues of the hessian $\nabla^2 \mathcal{L}(W)$.

Local optimization

Critical point condition:

$$\nabla \mathcal{L}(W) = 0.$$

Critical points are characterized by the eigenvalues of the hessian $\nabla^2 \mathcal{L}(W)$.
Let's for now pretend we are guaranteed to find a critical point with $\lambda_{min}(\nabla^2 \mathcal{L}(W)) \geq 0$.

How bad it is, compared to global minimum?

Local optimization

Critical point condition:

$$\nabla \mathcal{L}(W) = 0.$$

Critical points are characterized by the eigenvalues of the hessian $\nabla^2 \mathcal{L}(W)$.
Let's for now pretend we are guaranteed to find a critical point with $\lambda_{min}(\nabla^2 \mathcal{L}(W)) \geq 0$.

How bad it is, compared to global minimum?

Ideal result:
Every critical point W^* of $\mathcal{L}$ which is not a minimum has $\lambda_{min}(\nabla^2 \mathcal{L}(W^*)) < 0$, and every minimum of $\mathcal{L}$ is global.

Loss landscape

General feed-forward network:

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

$$\hat{y}(x, W) = W_H \sigma(W_{H-1} \dots \sigma(W_1 x) \dots).$$

Hard to analyze!

Loss landscape

General feed-forward network:

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

$$\hat{y}(x, W) = W_H \sigma(W_{H-1} \dots \sigma(W_1 x) \dots).$$

Hard to analyze!

Trivial special case: linear regression

$$\hat{y}(x, W) = Wx.$$

Loss landscape

General feed-forward network:

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

$$\hat{y}(x, W) = W_H \sigma(W_{H-1} \ldots \sigma(W_1 x) \ldots).$$

Hard to analyze!

Trivial special case: linear regression

$$\hat{y}(x, W) = Wx.$$

If $L(y, \hat{y})$ is convex wrt $\hat{y}$, then $\mathcal{L}(W)$ is convex wrt W.
Then we have a unique minimum! (and no saddles)

Loss landscape

From now on assume square loss:

$$L(y, \hat{y}) = \|y - \hat{y}\|_2^2.$$

More complex cases:

Loss landscape

From now on assume square loss:

$$L(y, \hat{y}) = \|y - \hat{y}\|_2^2.$$

More complex cases:

- One-layer non-linear net:

$$\hat{y}(x, W) = \sigma(Wx);$$

Loss landscape

From now on assume square loss:

$$L(y, \hat{y}) = \|y - \hat{y}\|_2^2.$$

More complex cases:

- One-layer non-linear net:

$$\hat{y}(x, W) = \sigma(Wx);$$

- Multi-layer linear net:

$$\hat{y}(x, W) = W_H W_{H-1} \dots W_1 x.$$

Which one is harder?

Linear nets

$$\mathcal{L}_{deep}(W) = \mathbb{E}_{x,y\sim\mathcal{D}}\|y - W_H W_{H-1} \dots W_1 x\|_2^2 \to \min_W .$$

Why complex?

Linear nets

$$\mathcal{L}_{deep}(W) = \mathbb{E}_{x,y\sim\mathcal{D}}\|y - W_H W_{H-1}\dots W_1 x\|_2^2 \to \min_W .$$

Why complex? Due to non-convexity!

Non-convex due to *weight-space symmetries:*

- If $(W_1, W_2, W_3, \dots, W_H)$ is a global minimum, then $(\alpha W_1, \alpha^{-1} W_2, W_3, \dots, W_H)$ is a global minimum too;
- $(\mathbf{0}, \dots, \mathbf{0})$ is a saddle point.

Linear nets

Consider finite dataset (X, Y);

Deep problem:

$$\mathcal{L}_{deep}(W) = \|Y - W_H W_{H-1} \dots W_1 X\|_F^2 \to \min_W .$$

Let d_i be the width of i-th layer; $X \in \mathbb{R}^{d_0 \times m}$, $Y \in \mathbb{R}^{d_H \times m}$.
$p := \arg\min_i d_i$ – bottleneck index.
Non-convex problem.

Linear nets

Consider finite dataset (X, Y);

Deep problem:

$$\mathcal{L}_{deep}(W) = \|Y - W_H W_{H-1} \dots W_1 X\|_F^2 \to \min_W .$$

Let d_i be the width of i-th layer; $X \in \mathbb{R}^{d_0 \times m}$, $Y \in \mathbb{R}^{d_H \times m}$.
$p := \arg\min_i d_i$ – bottleneck index.
Non-convex problem.

Shallow problem:

$$\mathcal{L}_{shallow}(R) = \|Y - RX\|_F^2 \to \min_R \quad s.t. \operatorname{rank}(R) \leq d_p.$$

Linear nets

Consider finite dataset (X, Y);

Deep problem:

$$\mathcal{L}_{deep}(W) = \|Y - W_H W_{H-1} \dots W_1 X\|_F^2 \to \min_W .$$

Let d_i be the width of i-th layer; $X \in \mathbb{R}^{d_0 \times m}$, $Y \in \mathbb{R}^{d_H \times m}$.
$p := \arg\min_i d_i$ – bottleneck index.
Non-convex problem.

Shallow problem:

$$\mathcal{L}_{shallow}(R) = \|Y - RX\|_F^2 \to \min_R \quad s.t. \operatorname{rank}(R) \leq d_p.$$

Non-convex problem too, unless $d_p = \min(d_0, d_H)$.

Linear nets

Consider finite dataset (X, Y);

Deep problem:

$$\mathcal{L}_{deep}(W) = \|Y - W_H W_{H-1} \dots W_1 X\|_F^2 \to \min_W .$$

Let d_i be the width of i-th layer; $X \in \mathbb{R}^{d_0 \times m}$, $Y \in \mathbb{R}^{d_H \times m}$.
$p := \arg\min_i d_i$ – bottleneck index.
Non-convex problem.

Shallow problem:

$$\mathcal{L}_{shallow}(R) = \|Y - RX\|_F^2 \to \min_R \quad s.t.\ \mathrm{rank}(R) \leq d_p.$$

Non-convex problem too, unless $d_p = \min(d_0, d_H)$.

Global minima are of the same value:

$$\min_W \mathcal{L}_{deep} = \min_R \mathcal{L}_{shallow} .$$

Linear nets

Consider finite dataset (X, Y);

Deep problem:

$$\mathcal{L}_{deep}(W) = \|Y - W_H W_{H-1} \dots W_1 X\|_F^2 \to \min_W .$$

Let d_i be the width of i-th layer; $X \in \mathbb{R}^{d_0 \times m}$, $Y \in \mathbb{R}^{d_H \times m}$.
$p := \arg\min_i d_i$ – bottleneck index.
Non-convex problem.

Shallow problem:

$$\mathcal{L}_{shallow}(R) = \|Y - RX\|_F^2 \to \min_R \quad s.t. \operatorname{rank}(R) \leq d_p.$$

Non-convex problem too, unless $d_p = \min(d_0, d_H)$.

Global minima are of the same value:

$$\min_W \mathcal{L}_{deep} = \min_R \mathcal{L}_{shallow}.$$

Are there any non-global minima of $\mathcal{L}_{deep}$?

Linear nets

Lu & Kawaguchi (2017)[4]:

Theorem 1:

If W is a local minimum of $\mathcal{L}_{deep}(W)$, than $R = W_H \dots W_1$ is a local minimum of $\mathcal{L}_{shallow}(R)$.

[4] https://arxiv.org/abs/1702.08580

[5] http://www.mit.edu/~kawaguch/publications/kawaguchi-nips16.pdf

Linear nets

Lu & Kawaguchi (2017)[4]:

Theorem 1:

If W is a local minimum of $\mathcal{L}_{deep}(W)$, than $R = W_H \dots W_1$ is a local minimum of $\mathcal{L}_{shallow}(R)$.

Theorem 2:

Every local minimum of $\mathcal{L}_{shallow}(R)$ is global.

[4] https://arxiv.org/abs/1702.08580

[5] http://www.mit.edu/~kawaguch/publications/kawaguchi-nips16.pdf

Linear nets

Lu & Kawaguchi (2017)[4]:

Theorem 1:

If W is a local minimum of $\mathcal{L}_{deep}(W)$, than $R = W_H \dots W_1$ is a local minimum of $\mathcal{L}_{shallow}(R)$.

Theorem 2:

Every local minimum of $\mathcal{L}_{shallow}(R)$ is global.

Corollary:

Every local minimum of $\mathcal{L}_{deep}(R)$ is global.

Almost the same result was obtained earlier in Kawaguchi (2016)[5].

[4] https://arxiv.org/abs/1702.08580

[5] http://www.mit.edu/~kawaguch/publications/kawaguchi-nips16.pdf

Linear nets

Deep linear net:

$$\mathcal{L}_{net}(W) = \|Y - W_H W_{H-1} \dots W_1 X\|_F^2 \to \min_W .$$

All local minima are global.

Linear nets

Deep linear net:

$$\mathcal{L}_{net}(W) = \|Y - W_H W_{H-1} \dots W_1 X\|_F^2 \to \min_W .$$

All local minima are global.

However, there exist "bad saddles" with no negative values of the Hessian (Kawaguchi, 2016), e.g.:

$\nabla \mathcal{L}_{net}(\mathbf{0}) = 0$ and $\nabla^2 \mathcal{L}_{net}(\mathbf{0}) = 0$ for $H \geq 3$.

Linear nets

Let $d_0 = d_1 = \ldots = d_H$;
Let $y = Rx + \xi$, where $\xi \sim \mathcal{N}(0, I)$.

Let's reparameterize our linear net as a ResNet:

$$\mathcal{L}_{resnet}(W) = \mathbb{E}_{x,\xi} \|y - (I + W_H)(I + W_{H-1}) \ldots (I + W_1)x\|_2^2 \to \min_W .$$

The same optimization problem! However,

[6] https://arxiv.org/abs/1611.04231

Linear nets

Let $d_0 = d_1 = \ldots = d_H$;
Let $y = Rx + \xi$, where $\xi \sim \mathcal{N}(0, I)$.

Let's reparameterize our linear net as a ResNet:

$$\mathcal{L}_{resnet}(W) = \mathbb{E}_{x,\xi} \|y - (I + W_H)(I + W_{H-1}) \ldots (I + W_1)x\|_2^2 \to \min_W .$$

The same optimization problem! However,

Theorem 1 (Hardt & Ma, 2016[6]):
Any critical point of $\mathcal{L}_{resnet}(W)$ for which $\max_{k=1,\ldots,H} \|W_k\| \leq \tau < 1$ is a global optimum.

[6] `https://arxiv.org/abs/1611.04231`

Linear nets

Let $d_0 = d_1 = \ldots = d_H$;
Let $y = Rx + \xi$, where $\xi \sim \mathcal{N}(0, I)$.

Let's reparameterize our linear net as a ResNet:

$$\mathcal{L}_{resnet}(W) = \mathbb{E}_{x,\xi} \| y - (I + W_H)(I + W_{H-1}) \ldots (I + W_1)x \|_2^2 \to \min_W .$$

The same optimization problem! However,

Theorem 1 (Hardt & Ma, 2016[6]):
Any critical point of $\mathcal{L}_{resnet}(W)$ for which $\max_{k=1,\ldots,H} \|W_k\| \le \tau < 1$ is a global optimum.

Theorem 2 (Hardt & Ma, 2016):
For large enough H there exists a global optimum with $\max_{k=1,\ldots,H} \|W_k\| < O(H^{-1})$.

[6] https://arxiv.org/abs/1611.04231