Theoretical assignment 2; 10 points total + 7 points extra

Theoretical Deep Learning course, MIPT

Problem 1

4 points.

Based on Hardt & Ma (2016)[1].

Let x and y be vectors of the same dimension d. Let R be the solution of a least square regression problem:

$$0 = \frac{\partial}{\partial R} \mathbb{E}_{x,y\sim\mathcal{D}} \|y - Rx\|_2^2, \tag{1}$$

where $\mathcal{D}$ denotes the data distribution. Consider loss of a linear ResNet:

$$\mathcal{L}(W_{1:H}) = \mathbb{E}_{x,y\sim\mathcal{D}} \|(I + W_H)\dots(I + W_1)x - y\|_2^2, \tag{2}$$

where all matrices W_k are square. Assume $\|W_k\|_2 < 1\ \forall k = 1, \dots, H$. Prove that

$$\frac{\partial \mathcal{L}}{\partial W_k} = 0 \quad \forall k = 1, \dots, H$$

is equivalent to

$$(I + W_H)\dots(I + W_1) = R. \tag{3}$$

Hence there are no local minima or saddle points in 1-vicinity of zero.

Note that we are not assuming $y = Rx + \xi$, where $\xi \sim \mathcal{N}(0, I_d)$; hence this result is a generalization of Theorem 2.2 of Hardt & Ma (2016). You can use any results that were actually proven in the paper.

Problem 2

7 points extra.

Based on Hardt & Ma (2016).

Assume $y = Rx + \xi$, where $\xi \sim \mathcal{N}(0, I_d)$, and R is a square matrix with $\det R > 0$.

Construct the solution $W_{1:H}$ of (3) such that for any $k = 1, \dots, H$ $\|W_k\|_2 \to 0$ as $H \to \infty$ (3 points extra).

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

You can use the following fact: for any orthogonal matrix U with determinant 1 we have $\|U^\alpha - I_d\|_2 \to 0$ as $\alpha \to 0$. You will receive up to 4 points extra for proving this fact.

Recall that solutions of (3) are exactly global minimizers of (2). Hence in this problem you are asked to construct a global minimizer with norm decaying to zero as number of layers grows. From this will follow that for sufficiently large H there exist a global minimum in 1-vicinity of zero.

We have already proven this statement for symmetric $R = U\Sigma U^T$ at the lecture (see also Section A.1 of the paper). In the paper you can find a proof for the general case; it is quite complicated, though. There is a simpler proof, very similar to symmetric case. Try to find it.

Problem 3

2 points.

Let $X \in \mathbb{R}^{d_0 \times m}$, $W \in \mathbb{R}^{d_1 \times d_0}$, where $d_0 < d_1 = m$ and all columns of X are distinct. Denote $G = WX \in \mathbb{R}^{d_1 \times m}$, $F = \sigma(G) \in \mathbb{R}^{d_1 \times m}$, where $\sigma(\cdot)$ is some non-linearity.

Note that G cannot be of full rank. However, we have proved (see lecture 4 or lemma 4 of Nguyen & Hein (2017)[2]) that for $\sigma(z) = (1 + \exp(-z))^{-1}$ the set of W for which F is not of full rank has Lebesgue measure zero. Does this result hold for ReLU, i.e. $\sigma(z) = \max(0, z)$?

Problem 4

4 points total.

Based on Kawaguchi & Kaelbling (2019)[3].

Consider an arbitrary model $f(x; \theta) \in \mathbb{R}$ differentiable wrt θ, a finite dataset $(x_i, y_i)_{i=1}^m$, where all $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$, and a convex (wrt $\hat{y}$) differentiable non-negative loss function $l(\hat{y}, y)$. Then, the loss of a model on a dataset is given as follows:

$$L(\theta) = \frac{1}{m} \sum_{i=1}^m l(f(x_i; \theta), y_i).$$

Assume $\min_\theta L(\theta) = 0$. Consider a modified loss of a model on a dataset:

$$\tilde{L}(\theta, w, a, b) = \frac{1}{m} \sum_{i=1}^m l(f(x_i; \theta) + a \exp(w^T x_i + b), y_i) + \lambda a^2,$$

where $w \in \mathbb{R}^d, a \in \mathbb{R}, b \in \mathbb{R}$ and $\lambda > 0$.

1. **2 points.** Consider $m = 1$. Prove that if (θ, w, a, b) is a local minimum of $\tilde{L}$, then θ is a global minimum of L.

[2] https://arxiv.org/abs/1704.08045
[3] https://arxiv.org/abs/1901.00279

2. **2 points.** This result looks strange: if we find a local minimum of $\tilde{L}$, then the corresponding θ will be a global minimum of L! We know, however, that global optimization is in general NP-complete. Try to explain for $m = 1$, where is the trick of this result. Is this result useful for optimization with gradient descent?