Gedächtnisprotokoll Klausur: Mathematics of Deep Learning 15.02.2020

90 Minuten. Keine Hilfsmittel, 50 Punkte, 1.0 bis 43 Punkte

Exercise 1 (6 Points)

Please decide, whether or not the following statements are correct. For every right decision you will earn one point. For every wrong decision, one point will be taken away. However, you are not able to get less than zero points and not more than six points for this task.

Statement	**True**	**False**
An advantage of training neural networks using gradient descent based methods, such as backpropagation, is that they can not get stuck in local minima of the loss function.		×
The double descent curve shows that over-parametrization is in general useful.	×	
A relevance map $(M_p)_p$ is conservative if $\prod_p M_p(x) = R_\varrho(\Phi)(x)$.		×
Continuous function on compact domain can be approximated well by neural network with three layers.	×	
In the direct inversion approach the neural network is only used for denoising.	×	
?		×

Exercise 2 (6 Points)

Consider the hypothesis space $\mathcal{H} \coloneqq \operatorname{span}\{\varphi_1, \varphi_2\}$, where $\varphi_1, \varphi_2 : \mathbb{R} \to \mathbb{R}$ are given as

$$\varphi_1(x) \coloneqq \sqrt{2}\,\mathbb{1}_{[1,2)}(x), \qquad \varphi_2(x) \coloneqq (1-\sqrt{2})x - 1 + 2\sqrt{2})\,\mathbb{1}_{[1,3]}(x).$$

Furthermore, let $S \coloneqq ((x_1, y_1), (x_2, y_2), (x_3, y_3)) \coloneqq ((1,1), (2,?), (3, 1.5))$ and $y \coloneqq (y_1, y_2, y_3)$. Compute the matrix $A \coloneqq (\varphi_j(x_i))_{i,j}$ and use it to compute the empirical target function $f_{\mathcal{H},S} \in \mathcal{H}$ which minimizes the empirical error $\mathcal{E}_S(f_{\mathcal{H},S})$.

Remark: You shall not compute the error $\mathcal{E}_S(f_{\mathcal{H}}, S)$.

We compute

$$A = \begin{pmatrix} \sqrt{2} & \sqrt{2} \\ 0 & 1 \\ 0 & 2-\sqrt{2} \end{pmatrix}$$

As the rows are linearly independent, the solution to $\arg\min_z \|Az-y\|$ is given by $w = A^+ y$, where $A^+ \coloneqq (A^T A)^{-1} A^T$ is the pseudoinverse of A. Tedious calculation yields

$$A^+ = \frac{1}{8\sqrt{2}-14} \begin{pmatrix} 8-7\sqrt{2} & 2 & 4-2\sqrt{2} \\ 0 & -2 & 2\sqrt{2}-4 \end{pmatrix}.$$

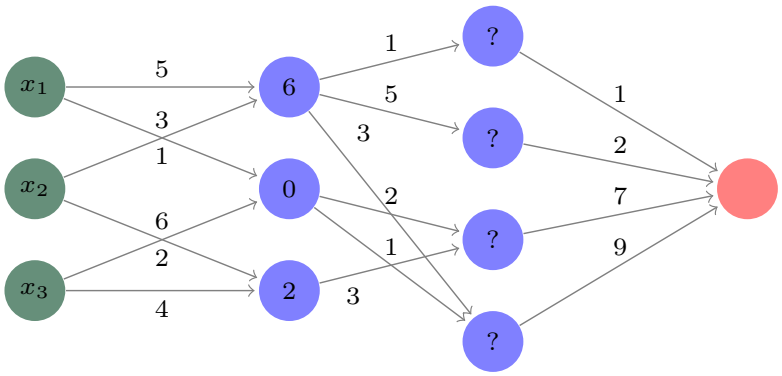
and therefore $f_{\mathcal{H},S} = w_1\varphi_1 + w_2\varphi_2$.

Exercise 3 (6 Points)

Consider the following neural network $\Phi \coloneqq ((A_1, b_1), (A_2, b_2), A_3, b_3))$, where

$$A_1 \coloneqq \begin{pmatrix} 5 & 1 & 0 \\ 3 & 0 & 6 \\ 0 & 2 & 4 \end{pmatrix}, \quad b_1 \coloneqq \begin{pmatrix} 6 \\ 0 \\ 2 \end{pmatrix}, \quad A_2 \coloneqq \begin{pmatrix} 1 & 0 & 0 \\ 5 & 0 & 0 \\ 0 & 2 & 3 \\ 3 & 1 & 0 \end{pmatrix}, \quad b_2 \coloneqq \begin{pmatrix} ? \\ ? \\ ? \\ ? \end{pmatrix}, \quad A_3 \coloneqq \begin{pmatrix} 1 \\ 2 \\ 7 \\ 9 \end{pmatrix}^\top$$

Visualize the architecture of Φ as a graph, which includes the weights and biases. Non-existent edges (caused by zero weights) shall not be visualized.



Exercise 4 (7 + 3 Points)

1. Let ϱ be the ReLU and

$$f : \mathbb{R}^2 \to \mathbb{R}, \; x \mapsto \begin{cases} -2, & \text{if } x_1 \geq 0, \\ 0, & \text{else.} \end{cases}$$

Construct a neural network Φ_ε with one hidden layer such that

(a) $|\mathcal{R}_\varrho(\Phi_\varepsilon)(x) - f(x)| \leq 2 \cdot \mathbb{1}_{[0,\varepsilon^2]}(x_1)$,

(b) $-2 \leq \Phi_\varepsilon(x) \leq 0$ for all $x \in \mathbb{R}^2$,

(c) $\|f - \mathcal{R}_\varrho(\Phi_\varepsilon)\|_{L^2([-1,1]^2)} \leq 2\varepsilon$.

2. For $a > 0$ and $g : \mathbb{R}^d \to \mathbb{R}$ define $T_a g(x) \coloneqq a g(ax)$. Furthermore, let Φ be a neural network such that $\mathcal{R}_\sigma(\Phi) : \mathbb{R}^d \to \mathbb{R}$. Construct a neural network $\tilde{\Phi}$ such that $\mathcal{R}_\sigma(\tilde{\Phi}) = T_a \mathcal{R}_\sigma(\Phi)$.

1. Choose $\Phi_\varepsilon \coloneqq ((A_1^\varepsilon, b_1, A_2, b_2))$, where

$$A_1^\varepsilon \coloneqq \begin{pmatrix} \varepsilon^{-2} & 0 \\ \varepsilon^{-2} & 0 \end{pmatrix}, \quad b_1 \coloneqq \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \qquad A_2 \coloneqq \begin{pmatrix} -2 \\ 2 \end{pmatrix}, \quad b_2 = 0.$$

Proof. We have

$$\mathcal{R}_\varrho(\Phi_\varepsilon(x)) = A_2 \varrho(A_1^\varepsilon x + b_1) = -2\varrho(\varepsilon^{-2} x_1) + 2\varrho(\varepsilon^{-2} x_1 - 1) =: R(x_1).$$

Let $x_1 \in [0, \varepsilon^2]$. Then we have $R(x_1) = -2\varepsilon^{-2} x_1$. Thus

$$|R(x_1) - f(x)| = \left| -2\varepsilon^{-2} x_1 - (-2) \right| = 2 \left| 1 - \varepsilon^{-2} x_1 \right| = 2(1 - \varepsilon^{-2} x_1) \leq 2.$$

For $x_1 < 0$ we have $R(x_1) = 0$ and for $x_1 > \varepsilon^2$ we have

$$R(x_1) = -2(\varepsilon^{-2} x_1) + 2(\varepsilon^{-2} x_1 - 1) = -2.$$

Finally with Fubinis theorem we have

$$\begin{aligned} \|f - \mathcal{R}_\varrho(\Phi_\varepsilon)\|^2_{L^2([-1,1]^2)} &= \int_{-1}^1 \int_{-1}^1 |f(x,y) - R(x)|^2 \,\mathrm{d}x \,\mathrm{d}y \\ &= 1 \cdot \int_{-1}^1 \left| -2 \cdot \mathbb{1}_{[0,\infty)}(x) - R(x) \right|^2 \mathrm{d}x \\ &\leq \int_{-1}^0 |0 - 0|^2 \,\mathrm{d}x + \int_0^{\varepsilon^2} |2|^2 \,\mathrm{d}x + \int_{\varepsilon^2}^1 (-2 - (-2))^2 \,\mathrm{d}x \\ &= 0 + \int_0^{\varepsilon^2} |2|^2 \,\mathrm{d}x + 0 = 4\varepsilon^2, \end{aligned}$$

implying $\|f - \mathcal{R}_\varrho(\Phi_\varepsilon)\|_{L^2([-1,1]^2)} \leq 2\varepsilon$. □

2. If $\Phi = ((A_k, b_k))_{k=1}^L$, then choose $\tilde{\Phi} \coloneqq ((\tilde{A}_k, \tilde{b}_k))_{k=1}^L$, where

$$\tilde{A}_k = a A_k, \ \tilde{b}_k = a b_k, \text{ for } k \in \{1, L\} \quad \text{and} \quad \tilde{A}_k = A_k, \ \tilde{b}_k = b_k \text{ for } k \in \{2, \dots, L-1\}.$$

Exercise 5 (3 + 7 Points)

1. Let X be a set and $H \subset \{h : X \to \{0,1\}\}$. Define the VC-dimension of H, VC-$\dim(H)$.
2. Consider $X \coloneqq [0,1]$ and $H \coloneqq \{\mathbb{1}_{[s,t]} \mid s, t \in [0,1]\}$. What is the VC-$\dim(H)$?

1. For $S \subset X$ the *restriction of H to S* is $H|_S \coloneqq \{h|_S : h \in H\}$. Then VC-$\dim(H) \coloneqq \sup\left\{m \in \mathbb{N} : \sup_{|S|\leq m} |H|_S| = 2^m\right\}$.

2. (a) Claim: VC-$\dim(H) \geq 2$.

   Proof. Let a and b be two points in $[0,1]$ with different x-coordinates $a_1 \neq b_1$ and the labels 0 and 1. Choose $s = a_1$ and $t = \frac{a_1+b_1}{2}$.

   If they are labelled differently the construction is analogous.

   (b) Claim: VC-$\dim(H) < 2$.

   Proof. Let $(x, y, z) \coloneqq ((0,0),(0,0.5),(0,1)) \subset [0,1]$ be three points with the labels $1, 0, 1$. For the points to be classified correctly we need the scattering function f to fulfill $f(0) = f(1) = 1$ but $f(0.5) = 0$, which is impossible as $f = \mathbb{1}_{[s,t]}$ for $s < t \in [0,1]$.

Exercise 6 (4 + 1 + 3 Points)

1. Which kinds of problems can be solved by the Douglas-Rachford-Algorithm? Explain the steps.
2. Which step can be improved by neural networks?
3. Describe one other instance where deep learning is used to solve inverse problems.

1. Explain inverse problems ...

   Douglas-Rachford: Define $\gamma > 0$ and the *proximal operator*

   $$\operatorname{prox}_f(v) \coloneqq \arg\min_z f(z) + \frac{1}{2}\|z - v\|^2$$

   in order to solve

   $$\min_{x\in\mathbb{R}^d} \|Ax - y\|^2 + \alpha\mathcal{R}(x)$$

   by the *"splitting" iteration*

   $$x_{k+1} \coloneqq \operatorname{prox}_{\gamma\alpha\mathcal{R}}(v_k), \qquad v_{k+1} \coloneqq v_k + \operatorname{prox}_{\gamma\|A\cdot - y\|^2}(2x_{k+1} - v_k) - x_{k+1}.$$

2. The proximal operator costly to compute analytically so we use a neural network in this step.
3. For example: MRI samples Radon transform, difficult if you can't have all angles, reconstruction / direct inversion with filtered backpropagation (FBP), train CNN to remove noise. Without taking FBP, CNN needs to learn physics of CT ...