NANYANG TECHNOLOGICAL UNIVERSITY

SEMESTER I TRIAL EXAMINATION

MH3520 – MATHEMATICS OF DEEP LEARNING

1. This examination paper contains **FIVE (5)** questions and comprises **FIVE (5)** printed pages.
2. Answer each question beginning on a **FRESH** page of the answer book.
3. This is a **RESTRICTED OPEN BOOK** exam. You are only allowed to bring in **ONE DOUBLE-SIDED A4-SIZE REFERENCE SHEET WITH TEXTS HANDWRITTEN OR TYPED ON THE A4 PAPER** (no sticky notes/post-it notes on the reference sheet).
4. Calculators may be used. However, you should write down systematically the steps in the workings.

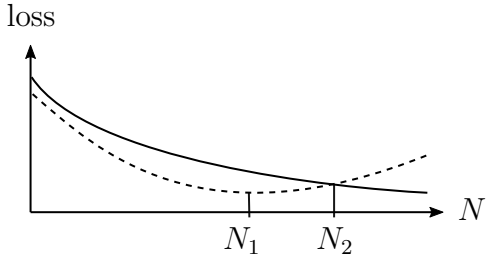
REMARK.
The questions in this trial exam are similar to the ones in the final exam. Hopefully, you will find them helpful in your preparation for the final exam.

NOTATION
The following notations and definitions are used throughout the exam paper.

- (x_i, y_i) are iid. random variables in $[-1, 1]^2$, indexed by $i \in \mathbb{N}$.
- $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is a smooth loss function.
- $\rho_1 : \mathbb{R} \to \mathbb{R}$ denotes the ReLU activation function $\rho_1(x) = \max\{0, x\}$.
- N-layer perception: $l_0 = x$, $l_n = \sigma(A_n l_{n-1} + b_n)$ for $n = 1, \cdots, N-1$, the output $l_N = A_N l_{N-1} + b_N$.

QUESTION 1: NUMERICAL OPTIMIZATION
Consider the task of training a multi-layer perceptron $h_w : \mathbb{R} \to \mathbb{R}$ with coefficients $w \in \mathbb{R}^N$ on a data-set $(x_i, y_i)_{i \in \{1,\ldots,n\}}$ with sample losses $f_i(w) := \ell(h_w(x_i), y_i)$ and learning rate $\alpha > 0$.

(a) Suppose that you train with various numbers N of network parameters, and you obtain the following training losses (solid) and validation losses (dashed):



Is N_1, N_2, or another value of N the optimal choice, and why?

> **Solution:** N_1 is optimal because it minimizes the validation loss. Of course, other considerations may have an influence, as well, including foremost computational time and memory requirements.

(b) Write down the updating rules for stochastic gradient descent without and with momentum.

See next page.

Solution: The updating rules for stochastic gradient are

$$w_{k+1} = w_k - \alpha \nabla f_k(w_k), \qquad w_{k+1} = w_k - \alpha \nabla f_k(w_k) + \beta(w_k - w_{k-1})$$

(c) Compute the gradients of the perceptron $h_w(x_i)$ and its loss $f_i(w)$ with respect to the weights w, for the specification

$$h_w(x) = \sum_{j=1}^{N} w_j \rho_1\Big(a_j x + b_j\Big), \qquad \ell(y, y') = (y - y')^2,$$

with input $x \in \mathbb{R}$, weights $w \in \mathbb{R}^N$, and untrainable parameters $a, b \in \mathbb{R}^N$.

Solution:

$$\nabla h_w(x_i) = \begin{pmatrix} \rho_1(a_1 x_i + b_1) \\ \vdots \\ \rho_1(a_N x_i + b_N) \end{pmatrix},$$

$$\nabla f_i(w) = 2\big(h_w(x_i) - y_i\big)\nabla h_w(x_i) = 2\big(h_w(x_i) - y_i\big) \begin{pmatrix} \rho_1(a_1 x_i + b_1) \\ \vdots \\ \rho_1(a_N x_i + b_N) \end{pmatrix}.$$

(d) In the situation of (c), with infinite sample size $n = \infty$, does the objective function (which one?) converge along the stochastic gradient descent? Do you expect to see linear or exponential convergence, and under what conditions?

Solution: The objective function is the expected loss

$$w \mapsto R(h_w) = E[f_i(w)],$$

where the expectation is with respect to the distribution of the data (x_i, y_i). The objective function is convex and quadratic and therefore satisfies the Łojasiewicz inequality. If the learning rate satisfies $\alpha < 1/L$, where L is the Lipschitz constant of $\nabla R(h_w)$, then the expected loss converges exponentially fast in L^1 along the stochastic gradient descent, alas not to the global minimum—this is

See next page.

due to the noise. To shut down the noise, the learning rate must be scheduled to decrease linearly in training time. For appropriate learning rate schedules, this leads to convergence in L^1 of the expected risk along the stochastic gradient descent to a global minimum. The minimum might be non-unique.

QUESTION 2: UNIVERSAL APPROXIMATION

(a) Are the following functions f perceptrons with the specified activation function ρ? Justify your answer.

$$f(x) = \max\{1 - |x|, 0\}, \qquad \rho(x) = \max\{x, 0\},$$
$$f(x) = x \qquad \rho(x) = \max\{x, 0\}^2.$$

Solution: First line: yes, because $f(x) = \rho(1 - \rho(x) - \rho(-x))$.
Second line: yes, because $2x = \rho(1 + x) + \rho(-1 - x) - \rho(x) - \rho(-x) - 1$.

(b) Let $H \subseteq C([-1, 1]^d)$ be the set of all multi-layer perceptrons with activation function $\rho_2 = \max\{0, x\}^2$. Use the Stone–Weierstrass theorem to show that H is dense in $C([-1, 1]^d)$.
Hint. Identity and multiplication can be realized as perceptrons with activation function ρ_2.

Solution:

- Multiplication of 2 variables can be realized by a network with 2 layers:

$$2xy = (x + y)_+^2 + (-x - y)_+^2 - (x)_+^2 - (-x)_+^2 - (y)_+^2 - (-y)_+^2.$$

- The identity can be represented by a network with 2 layers:

$$2x = (x + 1)_+^2 - (-x - 1)_+^2 - (x)_+^2 - 1.$$

- Thus, the multiplication of two perceptrons is again a perceptron.

See next page.

- It follows that H is an algebra.
- H is point-separating and contains all constant functions.
- Thus, the conditions of the Stone–Weierstrass theorem are satisfied, and H is dense in $C([-1,1]^d)$.

QUESTION 3: STATISTICAL LEARNING THEORY

(a) Let H be the set of 2-layer ReLU perceptrons with input dimension d, at most $N \in \mathbb{N}$ hidden neurons, and coefficients in $[-M, M]$, for some $M \in \mathbb{N}$. Show that H is compact in $C([-1,1]^d)$. Does H have finite covering numbers? Justify your answer.

Hint. Use that continuous images of compacts are compact.

Solution: The realization map is continuous. (This was a previous homework.) H is the continuous image of the compact set $[-M, M]^N$, hence compact. As $C([-1,1]^d)$ is a metric space, compactness implies total boundedness, which means that H has finite covering numbers.

(b) Let H be a subset of $X := L^2([-1,1]^d)$, and let $h_n, h^* \in H$ be minimizers of the empirical and expected risks R_n, R, respectively. Assume for some $\alpha, \sigma > 0$ that

$$R(h) = \sigma^2 + \|h - h^*\|^2, \qquad \sup_{h \in H} |R_n(h) - R(h)| \lesssim n^{-\alpha}.$$

Conclude that $\|h_n - h^*\| = O(n^{-\alpha/2})$.

Remark. Loosely speaking, thanks to the uniform convexity assumption on the expected risk, $\|h_n - h^*\|$ converges at half the convergence rate of the uniform generalization error.

Hint. Write $\|h_n - h\|^2$ as $R(h_n) - R(h)$ and use the bound on the uniform generalization error.

See next page.

Solution:

$$\begin{aligned} 0 \le \|h_n - h^*\|^2 &= R(h_n) - R(h^*) \\ &\le R(h_n) - R_n(h_n) + R_n(h_n) - R_n(h^*) + R_n(h^*) - R(h^*) \\ &\lesssim n^{-\alpha} + 0 + n^{-\alpha}. \end{aligned}$$

QUESTION 4: FUNCTION APPROXIMATION THEORY

(a) Consider the task of approximating the function $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^3$ in $C([-1, 1])$ by 3-layer networks with at most n non-zero weights at rate $O(n^{-\alpha})$. Show that every rate $\alpha < 2$ is achievable with activation function $\rho_1(x) := \max\{0, x\}$, and every rate $\alpha > 0$ is achievable with activation function $\rho_2(x) := \max\{0, x\}^2$.

Hint. For ρ_1, you may use dictionary learning with splines. For ρ_2, you may represent x^3 by networks of bounded size.

Solution: Let $X := C([-1, 1])$, and let ϕ denote the dictionary of piecewise linear splines in X on a dyadic grid, ordered by degree. Then, for any $\alpha < 2$,

$$f \in C^2([-1, 1]) \subseteq A^\alpha(X, \mathtt{H}(\phi)) \subseteq A^\alpha(X, \mathtt{W}(X, \rho_1, 2)) \subseteq A^\alpha(X, \mathtt{W}(X, \rho_1, 3)).$$

Here, the first inclusion is the direct estimate for spline approximation alluded to in the hint, the second inclusion is because $\mathtt{H}_n(\phi) \subseteq \mathtt{W}_n(X, \rho_1, 2)$, and the third inclusion is because the number of layers can be increased using the ReLU representation of the identity function.

Product of 4 variables can be realized as a network with three layers activation function ρ_2 and architecture $(1, 8, 4, 1)$. For instance, $x_1x_2 = \frac{1}{4}(\rho_2(x_1 + x_2) - \rho_2(x_1 - x_2) - \rho_2(-x_1 + x_2) + \rho_2(-x_1 - x_2))$ with 2 layer. Then FP to get x_1x_2 and x_3x_4, and non-sparse composition to get $x_1x_2x_3x_4$ with $2 + 2 - 1 = 3$ layers. Feed in the input $(x, x, x, 1)$ to this network realizes the function $f(x) = x^3$.

(b) Consider the task of approximating the function $f : \mathbb{R} \to \mathbb{R}$, $f(x) = \mathbb{1}_{(0,\infty)}$ by ReLU networks with at most n non-zero weights at rate $O(n^{-\alpha})$ in a function space X. Show for $X := L^\infty([-1, 1])$ that no positive rate α can be achieved, and show for

$X := L^2([-1,1])$ that every rate α can be achieved.

Hint. For $X := L^\infty([-1,1])$ use that ReLU networks are continuous, whereas the indicator function is not. For $X := L^2([-1,1])$, approximate the indicator function by 2-layer ReLU networks with 2 hidden neurons.

Solution: The indicator function has L^∞-distance at least $1/2$ to every continuous function and, in particular, to every ReLU perceptron.

By dominated convergence, the 2-layer networks $\epsilon^{-1}\rho_1(x+\epsilon)-\epsilon^{-1}\rho_1(x)$ converge in $L^2([-1,1])$ as $\epsilon \to 0$ to $\mathbb{1}_{(0,\infty)}$.

QUESTION 5: ROLE OF DEPTH

(a) Show that any L-layer ReLU perceptron $\mathbb{R} \to \mathbb{R}$ with N neurons per hidden layer is piecewise linear with at most $(2N)^{L-1}$ pieces.

Remark. This shows that the number of pieces is merely polynomial in N for fixed L, but exponential in L for fixed N.

Hint. Use induction by L. For the inductive step, write the perceptron with $L+1$ layers as $\sum_{i=1}^N a_i\rho(f_i(x))+b_i$, where each $f_i : \mathbb{R} \to \mathbb{R}$ is a perceptron with L layers.

Solution: The case $L = 1$ is clear. For the inductive step $L \to L+1$, consider $f = \sum_{i=1}^N a_i\rho(f_i(x))+b_i$ as in the hint. By the inductive hypothesis, each f_i is piecewise linear with at most $(2N)^{L-1}$ pieces. Therefore, $\rho \circ f_i$ is piecewise linear with at most $2 \cdot (2N)^{L-1}$ pieces. Therefore, f is piecewise linear with at most $2N \cdot (2N)^{L-1} = (2N)^L$ pieces.

(b) Show that the function $f(x) := \sin(8\pi x)^2$ can be approximated in $C([-1,1])$ at rate $n^{-\alpha}$ by ReLU perceptrons with at most n non-zero weights, for any (arbitrarily high) $\alpha > 0$.

Hint. The actual rate is much better, namely exponential.

Solution: Let $X := C([-1,1])$. Note that $2d + 10 = 12$. As the function f is real-analytic on $\mathbb{R}$, there exists $c > 0$ such that

$$E_{12\,n}(f, X, \mathtt{W}(X,\rho,\infty)) \le E_n(f, X, \mathtt{L}(X,\rho,12)) \lesssim e^{-cn^{1/3}} = O(n^{-\alpha}),$$

See next page.

for all $\alpha > 0$.

END OF PAPER