Problems and suggested solution
Group B

Question 1

1.MC1 [1 Point] Let $C([0,1]^n, \mathbb{R})$ denote the real valued continuous functions on the unit cube of $\mathbb{R}^n$. Let A be a subset of $C([0,1]^n, \mathbb{R})$. Under which conditions can we conclude that A is dense? (If there are more than one, choose the one with minimal assumptions)

(A) **TRUE:** A is a vector subspace additionally closed under multiplication, containing the constant function 1, which separates points.

(B) A is a vector subspace which additionally separates points.

(C) A is a vector subspace additionally closed under addition and multiplication.

(D) A contains all continuous functions that are linear combinations of constant functions and the identity function.

1.MC2 [1 Point] Let $\mathcal{NN}$ be the set of shallow neural networks in $C([0,1]^n, \mathbb{R})$ with respect to a non-constant bounded continuous activation function σ (understood as usual as restrictions of functions from $\mathbb{R} \to \mathbb{R}$). Which assertion is correct?

(A) $\mathcal{NN}$ is not a vector space but separates points.

(B) $\mathcal{NN}$ is a vector space closed under multiplication.

(C) **TRUE:** $\mathcal{NN}$ is a vector space additionally separating points.

(D) $\mathcal{NN}$ is a vector space but not separating points.

1.MC3 [1 Point] Let $\mathcal{NN}$ be the set of shallow neural networks in $C([a,b], \mathbb{R})$ with a non-constant bounded continuous activation function. The UAT tells that $\mathcal{NN}$ is dense. Let $\mathcal{L}$ be a point separating vector space of continuous, real valued functions on some compact space K. Denote by $\mathcal{NN}(\mathcal{L})$ the span of all compositions $f \circ l$ of a network $f \in \mathcal{NN}$ with a function $l \in \mathcal{L}$. Which assertion is correct?

(A) $\mathcal{NN}(\mathcal{L})$ is only dense in $C(K, \mathbb{R})$ if it is additionally closed under multiplication and point separating.

(B) $\mathcal{NN}(\mathcal{L})$ is not dense in $C(K, \mathbb{R})$.

(C) **TRUE:** $\mathcal{NN}(\mathcal{L})$ is only dense in $C(K, \mathbb{R})$ if it is additionally closed under multiplication.

(D) $\mathcal{NN}(\mathcal{L})$ is dense in $C(K, \mathbb{R})$.

1.MC4 [1 Point] Consider a shallow network with k hidden nodes on $[0,1]^d$ mapping to $\mathbb{R}$ (we always take a shift term here). Which assertion is correct?

(A) The number of parameters is kd for the affine function in the hidden layer and $k+1$ for the last layer.

(B) The number of parameters is $kd + k$ for the affine function in the hidden layer and k for the last layer.

(C) The number of parameters is kd for the affine function in the hidden layer and k for the last layer.

(D) **TRUE:** The number of parameters is $kd + k$ for the affine function in the hidden layer and $k + 1$ for the last layer.

1.MC5 [1 Point] Let $X \in C_0^1([0,1], \mathbb{R})$ s.t. $X_t = \sin(t)$ for all $t \in [0,1]$. Which is the correct signature $\mathrm{Sig}^{(2)}_{[0,1]}(X)$ up to second level?

(A) $\left(1, \cos(1), \int_0^1 t\cos(t)dt\right)$

(B) $\left(1, \sin(1), \int_0^1 t\cos(t)dt\right)$

(C) $\left(1, \cos(1), \int_0^1 \sin(t)\cos(t)dt\right)$

(D) **TRUE:** $\left(1, \sin(1), \int_0^1 \sin(t)\cos(t)dt\right)$

1.MC6 [1 Point] On the path space of bounded variation curves on $[0,T]$ starting at 0 with zeroth component equal to the third power of time, let A be the set of all possible linear combinations of signature components. Which assertion is correct?

(A) **TRUE:** A is a point separating vector space which is closed under multiplication.

(B) A is a vector space closed under multiplication but not point separating.

(C) A is a vector space but neither point separating nor closed under multiplication.

(D) A is a point separating vector space but not closed under multiplication.

1.MC7 [1 Point] Which of the following statements is NOT true for the signature map, the map from a curve to the collection of all its iterated integrals into the full sequence space on words?

(A) Both signature and truncated signature satisfies Chen's relations.

(B) The signature map is injective up to tree-like equivalences.

(C) The signature map is invariant under reparametrization.

(D) **TRUE:** The signature map is surjective, i.e. every sequence of numbers indexed by words is a signature of some bounded variation path.

1.MC8 [1 Point] What is the number of signature components up to depth M (M fold iterated integrals) of a curve $u : [0,T] \to \mathbb{R}^d$? (consider $d > 1$)

(A) **TRUE:** $\frac{d^{M+1}-1}{d-1}$

(B) $\frac{d^{M+1}-1}{d}$

(C) $\frac{(d+1)^M-1}{d-1}$

(D) $\frac{(d+1)^M-1}{d}$

1.MC9 [1 Point] Randomized neural networks, i.e. neural networks where some weights are chosen randomly and remain untrained, are surprisingly efficient in learning tasks. This is due to less training complexity and the fact that randomly chosen features often work very well. Which of the following statements about randomized networks is wrong?

(A) It depends on the nature of randomness and on the architecture whether the method works. Typically the last layer is trained. In any case some fine tuning of hyper-parameters is necessary.

(B) The method is often applied in case of recurrent neural networks and is called reservoir computing.

(C) **TRUE:** Even if all weights are chosen randomly (and no training is performed at all) there is a visible learning effect.

(D) Randomized neural networks are sometimes applied to achieve results in provable machine learning.

Question 2

2.MC1 [1 Point] Let ϕ be a self-financing portfolio in a financial market with bank account $S^0 = 1$ (zero interest rate) with value process V. Which of the following statements is correct for the value process V?

(A) $V_{t+1} - V_t = 0$ due to the self-financing condition for $t = 0, \dots, N-1$.

(B) $V_{t+1} - V_t = \sum_i \phi_t^i (S_{t+1}^i - S_t^i)$ for $t = 0, \dots, N-1$.

(C) The expression $V_{t+1} - V_t = \sum_i \phi_{t+1}^i S_{t+1}^i - \phi_t^i S_t^i$ cannot be simplified.

(D) **TRUE:** $V_{t+1} - V_t = \sum_i \phi_{t+1}^i (S_{t+1}^i - S_t^i)$ for $t = 0, \dots, N-1$.

2.MC2 [1 Point] Let S be a financial market with bank account $S^0 = 1$ (zero interest rate). Which of the following statements is correct?

(A) Absence of arbitrage can only be characterized by the existence of martingale measures and not by portfolio value processes.

(B) The market is free of arbitrage if for one self financing portfolio there is an equivalent martingale measure.

(C) The market is free of arbitrage if there is a self-financing portfolio, which looses, i.e. $V_0 = 0$ and $V_N \leq 0$ and $V_N \not\equiv 0$.

(D) **TRUE:** The market is free of arbitrage if there is no self-financing portfolio with $V_0 = 0$ and $V_N \geq 0$ and $V_N \not\equiv 0$.

2.MC3 [1 Point] Let $S = (S^0, \dots, S^d)$ be a financial market with bank account $S^0 = 1$ (zero interest rate) on a finite probability space, where the filtration is generated by the price process itself. How can we represent a generic $\mathcal{F}_N$ payoff by neural networks?

(A) We need to specify a stopping rule to fully determine the value of the derivative.

(B) **TRUE:** F is just a function of $S_0, \dots, S_N$ whose value is paid at time N.

(C) $F = (S_N - K)_+$.

(D) F is a function of S_N whose value is paid at time N.

2.MC4 [1 Point] Let S be a financial market with bank account $S^0 = 1$ (zero interest rate) on a finite probability space, where the filtration is generated by the price process itself. Consider now a derivative claim F of European type paying at time N. What does hedging mean?

(A) **TRUE:** Get as close as possible (under some to be specified criterion) to F with a self-financing portfolio, whose initial capital is defining a lower bound for the premium.

(B) Diversify the risk of F by investing in independent financial instruments.

(C) Find a self-financing portfolio which equals F at time N and has value 0 at time 0.

(D) Find a self-financing portfolio which is dominated by F at time N.

Question 3

In this series of questions we shall go through the main steps of a portfolio optimization problem solved by machine learning technology. Given a fixed stochastic process $(S_t)_{0\leq t\leq T}$ for the discounted price of one traded asset. We shall denote the wealth process by $V^\phi = V_0 + \int_0^\cdot \phi_{s-} dS_s$ of a self-financing portfolio holding ϕ_{t-} assets at price S_t at time t and with initial value V_0. Let u be a strictly concave utility function (e.g. logarithm or power utility). We aim to find an approximative solution for $\text{argsup}_\phi E[u(V_T^\phi)]$ for fixed initial wealth V_0 and a maturity time $T > 0$.

3.MC1 [1 Point] Assume that S is actually a martingale and that wealth processes are bounded from below. Does the problem have a simple solution?

(A) Yes, we buy one stock and wait until maturity.

(B) No, but the solution will be buy and hold of a non-trivial amount of stock.

(C) No, in all cases the solution is unknown and we have to run some training to learn the solution.

(D) **TRUE:** Yes, it is not worth investing in a martingale in case of a strictly concave utility function, i.e. $\phi = 0$.

3.MC2 [1 Point] Assume that we have a fixed grid of re-balancing dates $0 = t_0 < t_1 < \ldots < t_N = T$ and that there is a global bound on the amount of stocks held or sold (S is a general suitable stochastic process). Which of the following statements is true?

(A) It is sufficient to simulate trajectories of S as precise as possible on the given time grid of re-balancing dates to approximate wealth.

(B) There will be a closed formula for the wealth process given neural network strategies.

(C) One approximates V^ϕ also by a neural network to facilitate calculations.

(D) **TRUE:** If the strategy is only depending on the price path at the re-balancing grid points, then it is sufficient to simulate trajectories of S as precise as possible on the given time grid of re-balancing dates to approximate wealth.

3.MC3 [1 Point] We have to approximate the objective $E[u(V_T^\phi)]$ by a Monte Carlo Sampler. Does the quality of approximation really matter and how can we improve it?

(A) The training result does not depend on the quality of approximation since we do anyway apply sub-sampling.

(B) It is better to use analytic formulas for V^ϕ which are often available.

(C) We can improve it by reducing the variance of V_T^ϕ. This then helps to reduce the variance of $u(V^\phi)$ and therefore the quality of approximation.

(D) **TRUE:** We can improve it by using more samples of V_T^ϕ. If we do not use enough scenarios providing a good quality of approximation, overfitting can happen.

3.MC4 [1 Point] We parameterize the problem by neural networks. If there are no trading constraints and if strategies only depend on the price values at the re-balancing time grid before trading time, we can choose ...

(A) **TRUE:** ... a network of all the re-balancing time points before the trading time point to model ϕ at the time point.

(B) ... one network of all the re-balancing time points before the trading time point to model each ϕ_t.

(C) ... a network of each last re-balancing time point before the trading time point to model ϕ at the time point.

(D) ... one network of the last re-balancing time point before the trading time point to model each ϕ_t.

3.MC5 [1 Point] Having approximated the objective function with neural network strategies we can run a training algorithm to find the weights. Which of the following statements is true:

(A) Without proper explicit regularization the solution cannot be found by standard training technology.

(B) There will be a unique solution of the problem in terms of network parameters, which the algorithm will approximate quickly. This is due to the strict concavity of the problem.

(C) It is not enough to approximate the objective. We also have to approximate the gradients of the objective which is a completely different tasks.

(D) **TRUE:** The problem can be quite involved due to path dependence and having independent networks at each re-balancing time point. It might be that training takes a while.

Question 4

In this question, we look at time series forecasting using large language models (LLMs).

4.MC1 [1 Point] What does a generative pre-trained transformer do?

(A) A text is tokenized, i.e. decomposed into smaller units, and to each token the probability of appearance is calculated on training data.

(B) Language is always considered a Markov process on the set of tokens and the transition probabilities of token x_i to token x_j are calculated on training data.

(C) A text is tokenized, i.e. decomposed into smaller units, and to each context of fixed finite length of tokens $(x_1, \dots, x_N)$ the probability of appearance of $(x_1, \dots, x_N)$ is learned on training data.

(D) **TRUE:** A text is tokenized, i.e. decomposed into smaller units, and to each token x_{N+1} and a context of fixed finite length of tokens $(x_1, \dots, x_N)$ the probability of appearance of $(x_1, \dots, x_{N+1})$ is learned on training data.

4.MC2 [1 Point] In transformers, an important unit is called attention. It consists of queries $\mathbf{Q}[\mathbf{X}]$, keys $\mathbf{K}[\mathbf{X}]$, values $\mathbf{V}[\mathbf{X}]$ and attention weights $\mathbf{W}[\mathbf{X}]$. What is the correct way to calculate attention weights?

(A) **TRUE:** $\mathbf{W}[\mathbf{X}] = \text{Softmax}\left[\mathbf{K}[\mathbf{X}]^T\mathbf{Q}[\mathbf{X}]\right]$

(B) $\mathbf{W}[\mathbf{X}] = \text{Softmax}\left[\mathbf{K}[\mathbf{X}]^T\mathbf{V}[\mathbf{X}]\right]$

(C) $\mathbf{W}[\mathbf{X}] = \text{Tanh}\left[\mathbf{K}[\mathbf{X}]^T\mathbf{Q}[\mathbf{X}]\right]$

(D) $\mathbf{W}[\mathbf{X}] = \text{Tanh}\left[\mathbf{K}[\mathbf{X}]^T\mathbf{V}[\mathbf{X}]\right]$

4.MC3 [1 Point] In a transformer encoding block, what is the primary function of the positional encoding layer?

(A) To apply attention to the input sequence, focusing on the most relevant parts based on their position.

(B) To normalize the input sequence, ensuring that all elements have a similar magnitude.

(C) **TRUE:** To add positional information to the input sequence, preserving the original order.

(D) To project the input sequence into a higher-dimensional space, capturing complex relationships between elements.

4.MC4 [1 Point] In a transformer decoder, what is the primary purpose of masked attention when attending to the encoder's output?

(A) To project the query, key, and value vectors into a higher-dimensional space, allowing the decoder to capture more complex relationships between elements.

(B) To normalize the attention weights, preventing the decoder from focusing too much on any single token in the encoder's output.

(C) **TRUE:** To prevent the decoder from attending to future tokens in the input sequence, ensuring autoregressive generation.

(D) To ensure that the decoder only attends to the most relevant tokens in the encoder's output, based on their similarity to the current decoder state.

4.MC5 [1 Point] Why can the encoder compute in parallel while the decoder cannot?

(A) The decoder uses a different activation function that prevents parallel computation.

(B) **TRUE:** The decoder is inherently sequential due to the autoregressive nature of language generation.

(C) The encoder has more layers, which enables parallel computation.

(D) The encoder uses a different attention mechanism that allows for parallel computation.

4.MC6 [1 Point] Can we use LLM architectures for time series modeling?

(A) **TRUE:** Yes, if we tokenize the time series information appropriately, it is just a time series model with path dependent characteristics.

(B) No, time series models do never have path dependent characteristics.

(C) Yes, since LLMs give the correct output for any query.

(D) No, time series models do never have a discrete state space.