Model Predictive Control and Reinforcement Learning

– Fall School 2022 –

Prof. Joschka Bödecker, Prof. Moritz Diehl, Jasper Hoffmann, Florian Messerer

Surname: First Name: Matriculation number:

Pseudonym: Signature:

Each question is worth one point, and exactly one of the anwer possibilities will be correct.

1. Which statement about the Gauss-Newton method is wrong?

   (a) ☐   The QP solved at each iteration is always convex.

   (b) ☐   The Hessian approximation is computed from first-order derivatives only.

   (c) ☐   It can be applied to objective functions of arbitrary structure.

   (d) ☐   The convergence rate is in general linear.

   | 1 | |

2. Which statement about the `acados` software package is wrong?

   (a) ☐   `acados` implements an SQP method taylored to optimal control problems.

   (b) ☐   For integration, `acados` implements several implicit and explicit Runge-Kutta methods.

   (c) ☐   `acados` supports several QP solvers as its backend.

   (d) ☐   `acados` can solve Nonlinear Programs of general structure.

   | 1 | |

3. The Bellmann optimality equation can also be written as a... (choose the most specific correct option)

(a) ☐ Linear Program.	(b) ☐ Nonlinear Program.
(c) ☐ Quadratic Program.	(d) ☐ Semidefinite Program.

   | 1 | |

4. How many optimization variables does the NLP arising in the direct **single shooting** method have, if the system has n_x states, n_u controls, the initial value is fixed, and the time horizon is divided into N control intervals (piecewise-constant)?

(a) ☐ $(N+1)n_x + Nn_u$	(b) ☐ Nn_u	(c) ☐ $Nn_x^3 + Nn_u^2$	(d) ☐ $\frac{1}{3}N^3 n_u^3$

   | 1 | |

5. What is meant by monotonicity of DP? Formulate it using T as the DP operator, acting on value functions J and J'.

(a) ☐ $J' \geq J \Rightarrow TJ' \geq TJ$	(b) ☐ $TJ' \leq TJ \Rightarrow J' \leq J$
(c) ☐ $J' \leq J \Rightarrow TJ' \geq TJ$	(d) ☐ $TJ' \geq TJ \Rightarrow J' \leq J$

   | 1 | |

| Points on page (max. 5) | | |

6. Which statement about nonsmooth dynamical systems is wrong? Note that the systems are classified as systems with state-dependent switches (NSD2) and systems with state-dependent jumps (NSD3).

(a) ☐ For NSD2 systems, the accuracy of standard Runge-Kutta methods is always $O(h)$ irrespective of the integrator (polynomial) order (where h is the time-step).

(b) ☐ An NSD3 system can be transcribed into an NSD2 system via time-freezing.

(c) ☐ Most NSD2 systems can be described by a Filippov differential inclusion.

(d) ☐ When integrating an NSD2 system, it is not possible to exactly detect the switching time points.

| 1 | |

7. Which statement about Real Time Iterations (RTI) with Gauss-Newton Hessian approximation is wrong?

(a) ☐ RTI cannot handle state constraints.

(b) ☐ The computations can be split into a preparation and a feedback phase.

(c) ☐ After obtaining the most recent state estimate, only one QP is solved before returning a new control input.

(d) ☐ If the current state estimate remains constant, Gauss-Newton RTI is identical to a standard Gauss-Newton iteration.

| 1 | |

8. Which of the following statements on the sequential (single shooting) vs the simultaneous (multiple shooting) formulation of an Optimal Control Problem is wrong?

(a) ☐ The simultaneous formulation has an exploitable sparsity structure.

(b) ☐ The sequential formulation has less optimization variables.

(c) ☐ The simultaneous formulation is usually better at handling unstable nonlinear systems.

(d) ☐ The sequential formulation is always cheaper to solve.

| 1 | |

9. Which horizon is not included in the three horizons when we talk about "Three Horizon MPC"?

(a) ☐ Simulation horizon   (b) ☐ Optimization horizon

(c) ☐ Estimation horizon   (d) ☐ Missing horizon

| 1 | |

10. Which of the following statements about integrators is true?

(a) ☐ Implicit integrators cannot be a component of a Nonlinear Program (NLP).

(b) ☐ Implicit integators are better at handling stiff systems.

(c) ☐ Implicit integrators are always more accurate than explicit integrators.

(d) ☐ Implicit integrators are always computationally more efficient than explicit integrators.

| 1 | |

Points on page (max. 5) | |

11. Which of the following statements is correct?

   (a) ☐ Monte Carlo targets are biased but have high variance.

   (b) ☐ Temporal Difference methods use bootstrapping and thus targets are unbiased.

   (c) ☐ For a constant initial Q-value function, Q-learning targets for one fixed state have zero variance.

   (d) ☐ The concept of value functions as a mapping from states to $\mathbb{R}$, makes only sense with the Markov property of MDPs.

   | 1 |

12. What is importance sampling in the off-policy Monte Carlo methods not used for?

   (a) ☐ For variance reduction.

   (b) ☐ To get an unbiased target when the behaviour and target policy have different action distributions.

   (c) ☐ To allow for using an explorative behaviour policy while optimizing a greedy policy.

   (d) ☐ Training multiple policies with different objectives from a single interaction stream.

   | 1 |

13. To guarantee finding an optimal policy in a finite MDP we need to?

   (a) ☐ Use each action in every state at least once.　　(b) ☐ Explore every action at each state at least once.

   (c) ☐ Explore every state and action infinitely often.　　(d) ☐ Use each action of the action seed infinitely often.

   | 1 |

14. Which of the following statements is correct?

   (a) ☐ A dynamics model is easier to learn compared to a value function.

   (b) ☐ TD methods learn directly from raw experiences not necessarily needing a model.

   (c) ☐ Bootstrapping is used by TD methods in order to estimate the final outcome directly.

   (d) ☐ AlphaGo Zero uses a policy network for the rollout policy.

   | 1 |

15. Semi-gradient methods

   (a) ☐ are methods that only calculate a gradient regarding the prediction not the target.

   (b) ☐ are methods that only calculate a gradient regarding the target not the prediction.

   (c) ☐ are methods that calculate a gradient regarding the prediction and the target.

   (d) ☐ update a scalar parameter.

   | 1 |

Points on page (max. 5)

16. Imagine you want to apply the algorithms from this lecture on a real physical system. You get sensor input after each 0.05 seconds, but the execution of actions has a delay of 0.2 seconds. Is the Markov property fulfilled?

(a) ☐ Yes, if a history of the last 0.2 seconds is added to the state space.

(b) ☐ Only if a function approximator is used for the value function.

(c) ☐ Yes.

(d) ☐ No.

| 1 | |

17. Which is the formula of the UCB Bandit algorithm presented in the lecture:

(a) ☐ $\arg\max_a Q_t(a) + c\sqrt{\frac{N_t(a)}{\log t}}$.

(b) ☐ $\arg\max_a Q_t(a) + \frac{P(a)}{N+1}$.

(c) ☐ $\arg\max_a Q_t(a)$.

(d) ☐ $\arg\max_a Q_t(a) + c\sqrt{\frac{\log t}{N_t(a)}}$.

| 1 | |

18. What is true for Monte Carlo tree search:

(a) ☐ In general, the tree needs to cover each possible action to guarantee optimality.

(b) ☐ We count how often each action is taken only for the root node.

(c) ☐ It only does tree search before the game starts.

(d) ☐ At each step the nodes of the whole rollout are added to the tree.

| 1 | |

19. Which of the following components is part of AlphaGo Zero:

(a) ☐ A state value network v_θ to estimate the value of intermediate nodes.

(b) ☐ A visitation frequency network to estimate the visitation frequency.

(c) ☐ A state-action value network q_θ to estimate the value of actions at intermediate nodes.

(d) ☐ A policy-value network for an action prior probability and estimating the value of leaf nodes.

| 1 | |

20. The Deadly Triad of Reinforcement Learning is:

(a) ☐ Off-policy, function approximation, and bootstrapping.

(b) ☐ Off-policy, function approximation, and Monte Carlo.

(c) ☐ That the Bellman equation is not always true.

(d) ☐ On-policy, function approximation, and Monte Carlo.

| 1 | |

Points on page (max. 5) | |