Model Predictive Control and Reinforcement Learning

– Summer School 2021 –

Joschka Boedecker and Moritz Diehl

For the multiple choice questions, which give exactly one point, tick exactly one box for the right answer.

1. Which of the following functions $f(x)$, $f : \mathbb{R}^n \to \mathbb{R}$, is NOT convex ($c \in \mathbb{R}^n$, $A \in \mathbb{R}^{m\times n}$)?

(a) [x] $\|Ax\|_2^2 + \log(c^\top x)$	(b) [ ] $\|Ax\|_2^2 + \exp(c^\top x)$	(c) [ ] $c^\top x + \exp(\|Ax\|_2^2)$	(d) [ ] $-c^\top x + \|Ax\|_2^2$

1	

2. The local convergence rate of Newton's method is:

(a) [ ] superlinear	(b) [x] quadratic	(c) [ ] linear	(d) [ ] sublinear

1	

3. A point in the feasible set of an NLP that satisfies the KKT optimality conditions is

(a) [ ] the global minimum	(b) [ ] a boundary point
(c) [ ] a local minimum	(d) [x] a candidate for local minimum

1	

4. What is the most general (unconstrained) problem type to which the Gauss-Newton Hessian approximation is applicable?

(a) [x] non-linear least squares objective	(b) [ ] linear least-squares objective
(c) [ ] linear objective	(d) [ ] any convex objective

1	

5. How does CasADi compute derivatives?

(a) [ ] Finite differences	(b) [ ] Imaginary trick
(c) [x] Algorithmic Differentiation	(d) [ ] Symbolic Differentiation

1	

Points on page (max. 5)	

6. How many optimization variables does the NLP arising in the direct **multiple 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) [ ] Nn_u	(b) [ ] $Nn_x^3 + Nn_u^2$	(c) [x] $(N+1)n_x + Nn_u$	(d) [ ] $\frac{1}{3}N^3n_u^3$

1	

7. Regard an MPC optimization problem for $N = 10$ steps of the discrete time system $s^* = 2s + a$ with continuous state $s \in \mathbb{R}$ and continuous bounded control action $a \in [-1, 1]$. The stage cost is given by $c(s, a) = a^2$ and the terminal cost by $E(s) = 100s^2$. The initial state is $\bar{s}_0$. To which optimization problem class does the problem belong?

(a) [ ] Linear Programming (LP)	(b) [ ] Mixed Integer Programming (MIP) but not LP
(c) [x] Quadratic Programming (QP) but not LP	(d) [ ] Nonlinear Programming (NLP) but not QP

1	

8. Regard an MPC optimization problem for $N = 10$ steps of the discrete time system $s^+ = 2s^2 + a$ with continuous state $s \in \mathbb{R}$ and continuous bounded control action $a \in [-1, 1]$. The stage cost is given by $c(s, a) = a^2$ and the terminal cost by $E(s) = 100s^2$. The initial state is $\bar{s}_0$. To which optimization problem class does the problem belong?

(a) [ ] Linear Programming (LP)	(b) [ ] Mixed Integer Programming (MIP) but not LP
(c) [ ] Quadratic Programming (QP) but not LP	(d) [x] Nonlinear Programming (NLP) but not QP

1	

9. Regard dynamic programming for the discrete time system $s^+ = s + a$ with continuous state $s \in \mathbb{R}$ and continuous bounded control action $a \in [-1, 1]$, with zero stage cost $c(s, a) = 0$. We apply one step of dynamic programming (with operator T) to the value function $J_1(s) = \max(0, s)$. What is the resulting function $J_0 = TJ_1$?

(a) [x] $J_0(s) = \max(0, s-1)$	(b) [ ] $J_0(s) = \max(0, s+1)$
(c) [ ] $J_0(s) = 0$	(d) [ ] $J_0(s) = \max(s-1, s+1)$

1	

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

(a) [x] $J' \geq J \Rightarrow TJ' \geq TJ$	(b) [ ] $J' \leq J \Rightarrow TJ' \geq TJ$
(c) [ ] $TJ' \geq TJ \Rightarrow J' \leq J$	(d) [ ] $TJ' \leq TJ \Rightarrow J' \leq J$

1	

Points on page (max. 5)	

11. Which of the following components is *not* part of an MDP specification?

(a) [ ] Set of states	(b) [ ] Set of rewards	(c) [x] Policy	(d) [ ] Set of actions

1	

12. 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) [ ] No	(b) [ ] Yes	(c) [x] Yes, if a history of the last 0.2 seconds is added to the state space	(d) [ ] Only if a function approximator is used for the value function

1	

13. With what could you derive/calculate the value function $v_\pi(s)$ from the action-value function $q_\pi(s, a)$:

(a) [ ] With the Markov Decision Process (MDP)	(b) [x] With the policy π	(c) [ ] Not possible	(d) [ ] Only possible with both the MDP and the policy π

1	

14. Why is Q-learning an off-policy method?

(a) [ ] Because using an ϵ-greedy policy changes actions randomly	(b) [ ] Because Q-learning uses a bootstraped value, instead of a Monte-Carlo rollout
(c) [ ] Because we learn Q-values instead of a policy	(d) [x] Because we learn Q-values for the greedy policy, while using a different policy to interact with the environment.

1	

15. The correct Q-Learning update is:

(a) [ ] $Q(s, a) \leftarrow Q(s, a) + \alpha[r + \gamma \max_{a^\star} Q(s, a^\star) - Q(s', a^\star)]$	(b) [x] $Q(s, a) \leftarrow Q(s, a) + \alpha[r + \gamma \max_{a^\star} Q(s', a^\star) - Q(s, a)]$
(c) [ ] $Q(s, a) \leftarrow Q(s, a) + \alpha[r + \gamma Q(s', a) - Q(s', a)]$	(d) [ ] $Q(s, a) \leftarrow Q(s, a) + \alpha[r + \gamma Q(s', a) - Q(s, a)]$

1	

Points on page (max. 5)	

16. What can you control with the ϵ of an ϵ-greedy policy?

(a) ☐ The update size of a temporal difference method	(b) ☐ How much the agent emphasizes short term rewards vs long term rewards	(c) ☐ The randomness of the MDP	(d) ☒ How much the agent emphasizes exploration

1	

17. Target networks were introduced in order to:

(a) ☐ Introduce correlations into the sequence of observations	(b) ☐ Prevent forgetting past experiences	(c) ☒ Avoid oscillations during training which slow down learning	(d) ☐ Make RL problems less like Supervised Learning problems

1	

18. In policy gradient methods, what should a baseline ideally depend on?

(a) ☒ On the state	(b) ☐ On the action	(c) ☐ On state and action	(d) ☐ On nothing (i.e. it should be constant)

1	

19. Which of the following is true for Actor-Critic algorithms:

(a) ☐ Can be used only in problems with discrete actions	(b) ☒ They reduce gradient variance usually occurring in vanilla PG methods	(c) ☐ The usage of baselines is compulsory for variance reduction	(d) ☐ The actor learns a value function and the critic learns a policy

1	

20. We use experience replay to:

(a) ☐ Introduce correlations into the sequence of observations	(b) ☒ Prevent forgetting past experiences	(c) ☐ Avoid oscillations during the learning process	(d) ☐ Break the curse of dimensionality

1	

Points on page (max. 5)	

Empty page for calculations

Points on page (max. 0)		