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) $-c^T x + \|Ax\|_2^2$
   - (b) $c^T x + \exp(\|Ax\|_2^2)$
   - (c) $\|Ax\|_2^2 + \exp(c^T x)$
   - (d) $\|Ax\|_2^2 + \log(c^T x)$

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

   - (a) superlinear
   - (b) linear
   - (c) sublinear
   - (d) quadratic

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

   - (a) a candidate for local minimum
   - (b) the global minimum
   - (c) a boundary point
   - (d) a local minimum
4. What is the most general (unconstrained) problem type to which the Gauss-Newton Hessian approximation is applicable?

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

5. How does CasADi compute derivatives?

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

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

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 \( s_0 \). To which optimization problem class does the problem belong?

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

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 \( s_0 \). To which optimization problem class does the problem belong?

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

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) \( J_0(s) = 0 \)
(b) \( J_0(s) = \max(0, s + 1) \)
(c) \( J_0(s) = \max(s - 1, s + 1) \)
(d) \( J_0(s) = \max(0, s - 1) \)

10. What is meant by monotonicity of DP? Formulate it using \( T \) as the DP operator, acting on value functions \( J \) and \( J' \).
<table>
<thead>
<tr>
<th></th>
<th>(a) $TJ' \geq TJ \Rightarrow J' \leq J$</th>
<th>(b) $TJ' \leq TJ \Rightarrow J' \leq J$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(c) $J' \leq J \Rightarrow TJ' \geq TJ$</td>
<td>(d) $J' \geq J \Rightarrow TJ' \geq TJ$</td>
</tr>
</tbody>
</table>
11. Which of the following components is \textit{not} part of an MDP specification?

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Set of states</td>
<td>(b) Set of actions</td>
<td>(c) Set of rewards</td>
<td>(d) Policy</td>
</tr>
</tbody>
</table>

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

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 policy $\pi$  
(b) With the Markov Decision Process (MDP)  
(c) Only possible with both the MDP and the policy $\pi$  
(d) Not possible

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

(a) Because using an $\epsilon$-greedy policy changes actions randomly  
(b) Because we learn Q-values instead of a policy  
(c) Because Q-learning uses a bootstrapped value, instead of a Monte-Carlo rollout  
(d) Because we learn Q-values for the greedy policy, while using a different policy to interact with the environment.
15. The correct Q-Learning update is:

<table>
<thead>
<tr>
<th>Option</th>
<th>Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>( Q(s, a) \leftarrow Q(s, a) + \alpha[r + \gamma Q(s', a) - Q(s, a)] )</td>
</tr>
<tr>
<td>(b)</td>
<td>( Q(s, a) \leftarrow Q(s, a) + \alpha[r + \gamma Q(s', a) - Q(s, a)] )</td>
</tr>
<tr>
<td>(c)</td>
<td>( Q(s, a) \leftarrow Q(s, a) + \alpha[r + \gamma \max_{a^<em>} Q(s', a^</em>) - Q(s', a^*)] )</td>
</tr>
<tr>
<td>(d)</td>
<td>( Q(s, a) \leftarrow Q(s, a) + \alpha[r + \gamma \max_{a^<em>} Q(s', a^</em>) - Q(s, a)] )</td>
</tr>
</tbody>
</table>
16. What can you control with the $\epsilon$ of an $\epsilon$-greedy policy?

<table>
<thead>
<tr>
<th>(a) The randomness of the MDP</th>
<th>(b) How much the agent emphasizes exploration</th>
<th>(c) How much the agent emphasizes short term rewards vs long term rewards</th>
<th>(d) The update size of a temporal difference method</th>
</tr>
</thead>
<tbody>
<tr>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
<td>[ ]</td>
</tr>
</tbody>
</table>

1 point
17. Target networks were introduced in order to:

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

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

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

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

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

20. We use experience replay to:

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

1
Empty page for calculations