Список вопросов к экзамену по курсу “Обучение с подкреплением”, осень 2021

1. Кросс-энтропийный метод в общем виде. Его применение для решения задач оптимизации и задач обучения с подкреплением.
2. Уравнения Беллмана для функций ценности. Алгоритмы Policy/Value Iteration.
3. Табличные методы: Монте-Карло, Q-learning, Sarsa.
4. Алгоритм DQN и его модификации: Double DQN, приоритезированный буфер, дуэльная архитектура, шумные сети, многошаговый DQN, память.
5. Distributional-подход в RL. Алгоритмы c51 и QR-DQN.
6. Подход Policy gradient. Алгоритмы Reinforce и A2C.
7. Метод Trust-Region Policy Optimization (TRPO), его теоретическое обоснование.
8. Bias-variance trade-off в обучении с подкреплением. Оценка GAE. Алгоритм Proximal Policy Optimization (PPO).
9. Детерминированный градиент по политике. Off-policy алгоритмы для задач непрерывного управления: DDPG, Twin Delayed DDPG (TD3)
10. Обучение с подкреплением с добавлением энтропии. Алгоритм Soft Actor-Critic.
11. Имитационное обучение и обратное обучение с подкреплением. Схема Guided Cost Learning. Генеративно-состязательное имитационное обучение (GAIL)
12. Задача многоруких бандитов, UCB-бандиты. Внутренняя мотивация: дистилляция случайной сети (RND) и внутренний модуль любопытства (ICM).
13. Monte Carlo Tree Search в общем виде. Методы AlphaZero и MuZero.
14. Линейно-квадратичный регулятор и его итеративная версия. Общая схема Model-based RL.

Теоретический минимум

Вопросы из этой части охватывают базовые математические понятия и алгоритмы, которые активно используются в курсе. Незнание ответа на любой вопрос из данной части автоматически влечёт за собой неудовлетворительную оценку по экзамену.

1. Постановка задачи обучения с подкреплением в виде марковского процесса принятия решений. Примеры прикладных задач.
2. Уравнения Беллмана для функций ценности. Связь функций ценности между собой.
3. Дилемма exploration vs. exploitation. Подходы к её решению.
4. Понятие on-policy и off-policy алгоритмов обучения с подкреплением. Примеры.
5. Виды RL алгоритмов: model-based, model-free (value-based, policy gradient, эволюционный подходы).
6. Схема метода Q-learning и DQN.
7. Схема метода A2C.
8. Понятия soft Q-функции и V-функции. Уравнения Беллмана для них. Оптимальная политика для них.
9. Схема метода UCB-бандит.