Deep Reinforcement Learning Algorithms: Group Presentations

Topics in RL: S25

1 Project Presentation Guidelines

All project presentations must adhere to the following guidelines:

- Presentations must be created using LaTeX Beamer. Other formats will not be accepted.
- Each team member must contribute to the preparation of the slides.
- The slides must explicitly indicate which team member contributed to which portion.
- Algorithms presented in the papers must be clearly explained with appropriate equations and mathematical derivations.
- Pasting equations as images is strictly not allowed. All mathematical expressions must be written using LaTeX math mode.
- The presentation should be cohesive, well-structured, and logically organized. Marks will be awarded based on how effectively the presentation conveys the core idea of the paper.
- A proof outline for convergence must be provided if it's there in the paper.
- The final section should include 2-3 slides suggesting new innovative ideas or potential research directions that extend the existing algorithms.

2 Group-wise Algorithm Progression

2.1 Group 5: Value-based Methods

- **DQN (Deep Q-Network) [14]**: Introduces deep learning to Q-learning, using a neural network to approximate the Q-function.

- **Double DQN [22]**: Addresses overestimation bias in Q-learning by decoupling action selection and evaluation.
- **Dueling DQN [24]**: Introduces separate streams for state-value and advantage estimation, improving learning efficiency.
- **Soft Q-Learning [7]**: Learns maximum entropy policies using an energy-based formulation and a Boltzmann distribution for optimal policy representation.

2.2 Group 1: Policy Gradient Methods

- **Natural Policy Gradient (NPG) [1]**: NPG uses a Fisher information matrix to scale gradient updates, leading to more stable learning.
- **Generalized Advantage Estimation (GAE) [18]**: GAE reduces variance in policy gradient updates by introducing a more refined advantage function.
- **New perspectives of Natural Policy Gradient [12]**: A second-order optimization method using the Fisher information matrix as a Hessian substitute, improving convergence speed and robustness through techniques like trust regions and Tikhonov regularization. Group is expected to present Geometric interpretation of NPG from this paper.

2.3 Group 7: Deterministic Policy Gradient Methods and Goal Relabelling

- **DDPG (Deep Deterministic Policy Gradient) [11]**: Combines DPG with deep function approximation, using experience replay and target networks.
- **TD3 (Twin Delayed DDPG) [6]**: Addresses overestimation bias in DDPG by introducing twin Q-networks and target smoothing.
- **HER (Hindsight Experience Replay) [2]**: Enables learning from failed episodes by relabelling goals in hindsight, making RL feasible for sparse-reward environments.

2.4 Group 8: Actor-Critic Methods

- **Naive Actor-Critic [4]**: Combines policy gradient and value function approximation for more stable learning.
- **A2C (Advantage Actor-Critic) [13]**: Synchronously updates multiple workers to improve efficiency.
- **A3C (Asynchronous Advantage Actor-Critic) [13]**: Introduces asynchronous updates, improving exploration and stability.

- **ACER (Actor-Critic with Experience Replay)** [**23**]: Introduces experience replay and trust region updates in an actor-critic framework.

2.5 Group 3: Trust Region Policy Gradient Approaches

- **TRPO (Trust-Region Policy Optimization)** [**17**]: Enforces a constraint on policy updates to ensure stability.
- **PPO (Proximal Policy Optimization)** [**19**]: Simplifies TRPO by using a clipped objective, making it more practical and efficient.
- **GRPO (Group Relative Policy Optimization)** [**20**]: Eliminates the need for an explicit value function by leveraging relative rewards within a sampled group of outputs, improving variance reduction and aligning well with the comparative nature of reward models.

2.6 Group 4: Soft Actor-Critic and its variants

- **SAC (Soft Actor-Critic)** [**8**]: Optimizes policies using entropy maximization, leading to better exploration.
- **SAC with Adjustable Temperature** [**9**]: Extends SAC by dynamically adjusting temperature, balancing exploration and exploitation.
- **SACHER (Soft Actor-Critic with Hindsight Experience Replay)** [**10**]: Enhances SAC by incorporating HER [2], enabling learning from both successful and failed attempts, improving sample efficiency in goal-conditioned tasks such as UAV navigation.

2.7 Group 6: Model based RL

- **Dyna-Q and Dyna-Q++** [**21**]: Model-based reinforcement learning methods that integrate planning, acting, and learning by using a learned model to generate additional training experiences, improving sample efficiency and adaptation.
- **MuZero** [16]: Achieves superhuman performance in complex domains by combining tree-based planning with a learned model that predicts policy, value, and reward, without requiring explicit environment dynamics.
- **Stochastic MuZero** [**3**]: Extends MuZero to learn and plan with stochastic models, improving performance in partially observed and inherently stochastic environments, such as 2048 and backgammon, while maintaining MuZero's success in structured games like Go.

2.8 Group 2: Imitation learning

- **DAgger (Dataset Aggregation) [15]**: Iteratively improves a policy by querying an expert when encountering unseen states, mitigating distributional shift.
- **MaxEnt IRL (Maximum Entropy Inverse Reinforcement Learning) [26]**: An alternative to behavioral cloning is to infer the expert's underlying reward function R which is referred to as Inverse RL. MaxEnt IRL Learns a reward function by matching expert feature expectations while ensuring minimal unintended path preferences via entropy regularization.
- **Action Chunking with Transformers (ACT)** [25]: Enables low-cost robots to perform fine manipulation tasks using end-to-end imitation learning from real demonstrations, addressing compounding errors with action chunking.
- **Diffusion Policy** [5]: Frames visuomotor policy learning as a conditional denoising diffusion process, achieving state-of-the-art performance across multiple robotic manipulation benchmarks.

  Note: The group is free to choose between (ACT) [25] and Diffusion Policy [5] for their presentations but first two are mandatory.

References

[1] Shun-ichi Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

[2] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. *Advances in neural information processing systems*, 30, 2017.

[3] Ioannis Antonoglou, Julian Schrittwieser, Sherjil Ozair, Thomas K Hubert, and David Silver. Planning in stochastic environments with a learned model. In *International Conference on Learning Representations*, 2021.

[4] Andrew G Barto, Richard S Sutton, and Charles W Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE transactions on systems, man, and cybernetics*, (5):834–846, 1983.

[5] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, page 02783649241273668, 2023.

[6] Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pages 1587–1596. PMLR, 2018.

[7] Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *International conference on machine learning*, pages 1352–1361. PMLR, 2017.

[8] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. Pmlr, 2018.

[9] Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018.

[10] Myoung Hoon Lee and Jun Moon. Deep reinforcement learning-based model-free path planning and collision avoidance for uavs: A soft actor–critic with hindsight experience replay approach. *ICT Express*, 9(3):403–408, 2023.

[11] Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

[12] James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020.

[13] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937. PmLR, 2016.

[14] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[15] Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635. JMLR Workshop and Conference Proceedings, 2011.

[16] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart,

Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.

[17] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

[18] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

[19] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[20] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[21] Richard S Sutton. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Machine learning proceedings 1990*, pages 216–224. Elsevier, 1990.

[22] Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

[23] Ziyu Wang, Victor Bapst, Nicolas Heess, Volodymyr Mnih, Remi Munos, Koray Kavukcuoglu, and Nando De Freitas. Sample efficient actor-critic with experience replay. *arXiv preprint arXiv:1611.01224*, 2016.

[24] Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Hasselt, Marc Lanctot, and Nando Freitas. Dueling network architectures for deep reinforcement learning. In *International conference on machine learning*, pages 1995–2003. PMLR, 2016.

[25] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023.

[26] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.