Key Papers in Deep RL

What follows is a list of papers in deep RL that are worth reading. This is *far* from comprehensive, but should provide a useful starting point for someone looking to do research in the field.

Table of Contents

1. Model-Free RL

a. Deep Q-Learning

[1] Playing Atari with Deep Reinforcement Learning, Mnih et al, 2013. **Algorithm: DQN.**

[2] Deep Recurrent Q-Learning for Partially Observable MDPs, Hausknecht and Stone, 2015. **Algorithm: Deep Recurrent Q-Learning.**

[3] Dueling Network Architectures for Deep Reinforcement Learning, Wang et al, 2015. **Algorithm: Dueling DQN.**

[4] Deep Reinforcement Learning with Double Q-learning, Hasselt et al 2015. **Algorithm: Double DQN.**

[5] Prioritized Experience Replay, Schaul et al, 2015. **Algorithm: Prioritized Experience Replay (PER).**

[6] Rainbow: Combining Improvements in Deep Reinforcement Learning, Hessel et al, 2017. **Algorithm: Rainbow DQN.**

b. Policy Gradients

[7] Asynchronous Methods for Deep Reinforcement Learning, Mnih et al, 2016. **Algorithm: A3C.**

[8] Trust Region Policy Optimization, Schulman et al, 2015. **Algorithm: TRPO.**

[9] High-Dimensional Continuous Control Using Generalized Advantage Estimation, Schulman et al, 2015. **Algorithm: GAE.**

[10] Proximal Policy Optimization Algorithms, Schulman et al, 2017. **Algorithm: PPO-Clip, PPO-Penalty.**

[11] Emergence of Locomotion Behaviours in Rich Environments, Heess et al, 2017. **Algorithm: PPO-Penalty.**

[12] Scalable trust-region method for deep reinforcement learning using Kronecker-factored approximation, Wu et al, 2017. **Algorithm: ACKTR.**

[13] Sample Efficient Actor-Critic with Experience Replay, Wang et al, 2016. **Algorithm: ACER.**

[14] Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor, Haarnoja et al, 2018. **Algorithm: SAC.**

c. Deterministic Policy Gradients

[15] Deterministic Policy Gradient Algorithms, Silver et al, 2014. **Algorithm: DPG.**

[16] Continuous Control With Deep Reinforcement Learning, Lillicrap et al, 2015. **Algorithm: DDPG.**

[17] Addressing Function Approximation Error in Actor-Critic Methods, Fujimoto et al, 2018. **Algorithm: TD3.**

d. Distributional RL

[18] A Distributional Perspective on Reinforcement Learning, Bellemare et al, 2017. **Algorithm: C51.**

[19] Distributional Reinforcement Learning with Quantile Regression, Dabney et al, 2017. **Algorithm: QR-DQN.**

[20] Implicit Quantile Networks for Distributional Reinforcement Learning, Dabney et al, 2018. **Algorithm: IQN.**

[21] Dopamine: A Research Framework for Deep Reinforcement Learning, Anonymous, 2018. **Contribution:** Introduces Dopamine, a code repository containing implementations of DQN, C51, IQN, and Rainbow. Code link.

e. Policy Gradients with Action-Dependent Baselines

[22] Q-Prop: Sample-Efficient Policy Gradient with An Off-Policy Critic, Gu et al, 2016. **Algorithm: Q-Prop.**

[23] Action-depedent Control Variates for Policy Optimization via Stein's Identity, Liu et al, 2017. **Algorithm: Stein Control Variates.**

[24] The Mirage of Action-Dependent Baselines in Reinforcement Learning, Tucker et al, 2018. **Contribution:** interestingly, critiques and reevaluates claims from earlier papers (including Q-Prop and stein control variates) and finds important methodological errors in them.

f. Path-Consistency Learning

[25] Bridging the Gap Between Value and Policy Based Reinforcement Learning, Nachum et al, 2017. **Algorithm: PCL.**

[26] Trust-PCL: An Off-Policy Trust Region Method for Continuous Control, Nachum et al, 2017. **Algorithm: Trust-PCL.**

g. Other Directions for Combining Policy-Learning and Q-Learning

[27] Combining Policy Gradient and Q-learning, O'Donoghue et al, 2016. **Algorithm: PGQL.**

[28] The Reactor: A Fast and Sample-Efficient Actor-Critic Agent for Reinforcement Learning, Gruslys et al, 2017. **Algorithm: Reactor.**

[29] Interpolated Policy Gradient: Merging On-Policy and Off-Policy Gradient Estimation for Deep Reinforcement Learning, Gu et al, 2017. **Algorithm: IPG.**

[30] Equivalence Between Policy Gradients and Soft Q-Learning, Schulman et al, 2017. **Contribution:** Reveals a theoretical link between these two families of RL algorithms.

h. Evolutionary Algorithms

[31] Evolution Strategies as a Scalable Alternative to Reinforcement Learning, Salimans et al, 2017. **Algorithm: ES.**

2. Exploration

a. Intrinsic Motivation

[32] VIME: Variational Information Maximizing Exploration, Houthooft et al, 2016. **Algorithm: VIME.**

[33] Unifying Count-Based Exploration and Intrinsic Motivation, Bellemare et al, 2016. **Algorithm: CTS-based Pseudocounts.**

[34] Count-Based Exploration with Neural Density Models, Ostrovski et al, 2017. **Algorithm: PixelCNN-based Pseudocounts.**

[35] #Exploration: A Study of Count-Based Exploration for Deep Reinforcement Learning, Tang et al, 2016. **Algorithm: Hash-based Counts.**

[36] EX2: Exploration with Exemplar Models for Deep Reinforcement Learning, Fu et al, 2017. **Algorithm: EX2.**

[37] Curiosity-driven Exploration by Self-supervised Prediction, Pathak et al, 2017. **Algorithm: Intrinsic Curiosity Module (ICM).**

[38] Large-Scale Study of Curiosity-Driven Learning, Burda et al, 2018. **Contribution:** Systematic analysis of how surprisal-based intrinsic motivation performs in a wide variety of environments.

[39] Exploration by Random Network Distillation, Burda et al, 2018. **Algorithm: RND.**

b. Unsupervised RL

[40] Variational Intrinsic Control, Gregor et al, 2016. **Algorithm: VIC.**

[41] Diversity is All You Need: Learning Skills without a Reward Function, Eysenbach et al, 2018. **Algorithm: DIAYN.**

[42] Variational Option Discovery Algorithms, Achiam et al, 2018. **Algorithm: VALOR.**

3. Transfer and Multitask RL

[43] Progressive Neural Networks, Rusu et al, 2016. **Algorithm: Progressive Networks.**

[44] Universal Value Function Approximators, Schaul et al, 2015. **Algorithm: UVFA.**

[45] Reinforcement Learning with Unsupervised Auxiliary Tasks, Jaderberg et al, 2016. **Algorithm: UNREAL.**

[46] The Intentional Unintentional Agent: Learning to Solve Many Continuous Control Tasks Simultaneously, Cabi et al, 2017. **Algorithm: IU Agent.**

[47] PathNet: Evolution Channels Gradient Descent in Super Neural Networks, Fernando et al, 2017. **Algorithm: PathNet.**

[48] Mutual Alignment Transfer Learning, Wulfmeier et al, 2017. **Algorithm: MATL.**

[49] Learning an Embedding Space for Transferable Robot Skills, Hausman et al, 2018.

[50] Hindsight Experience Replay, Andrychowicz et al, 2017. **Algorithm: Hindsight Experience Replay (HER).**

4. Hierarchy

[51] Strategic Attentive Writer for Learning Macro-Actions, Vezhnevets et al, 2016. **Algorithm: STRAW.**

[52] FeUdal Networks for Hierarchical Reinforcement Learning, Vezhnevets et al, 2017. **Algorithm: Feudal Networks**

[53] Data-Efficient Hierarchical Reinforcement Learning, Nachum et al, 2018. **Algorithm: HIRO.**

5. Memory

[54] Model-Free Episodic Control, Blundell et al, 2016. **Algorithm: MFEC.**

[55] Neural Episodic Control, Pritzel et al, 2017. **Algorithm: NEC.**

[56] Neural Map: Structured Memory for Deep Reinforcement Learning, Parisotto and Salakhutdinov, 2017. **Algorithm: Neural Map.**

[57] Unsupervised Predictive Memory in a Goal-Directed Agent, Wayne et al, 2018. **Algorithm: MERLIN.**

[58] Relational Recurrent Neural Networks, Santoro et al, 2018. **Algorithm: RMC.**

6. Model-Based RL

a. Model is Learned

[59] Imagination-Augmented Agents for Deep Reinforcement Learning, Weber et al, 2017. **Algorithm: I2A.**

[60] Neural Network Dynamics for Model-Based Deep Reinforcement Learning with Model-Free Fine-Tuning, Nagabandi et al, 2017. **Algorithm: MBMF.**

[61] Model-Based Value Expansion for Efficient Model-Free Reinforcement Learning, Feinberg et al, 2018. **Algorithm: MVE.**

[62] Sample-Efficient Reinforcement Learning with Stochastic Ensemble Value Expansion, Buckman et al, 2018. **Algorithm: STEVE.**

[63] Model-Ensemble Trust-Region Policy Optimization, Kurutach et al, 2018. **Algorithm: ME-TRPO.**

[64] Model-Based Reinforcement Learning via Meta-Policy Optimization, Clavera et al, 2018. **Algorithm: MB-MPO.**

[65] Recurrent World Models Facilitate Policy Evolution, Ha and Schmidhuber, 2018.

b. Model is Given

[66] Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm, Silver et al, 2017. **Algorithm: AlphaZero.**

[67] Thinking Fast and Slow with Deep Learning and Tree Search, Anthony et al, 2017. **Algorithm: ExIt.**

7. Meta-RL

[68] RL^2: Fast Reinforcement Learning via Slow Reinforcement Learning, Duan et al, 2016. **Algorithm: RL^2.**

[69] Learning to Reinforcement Learn, Wang et al, 2016.

[70] Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks, Finn et al, 2017. **Algorithm: MAML.**

[71] A Simple Neural Attentive Meta-Learner, Mishra et al, 2018. **Algorithm: SNAIL.**

8. Scaling RL

[72] Accelerated Methods for Deep Reinforcement Learning, Stooke and Abbeel, 2018. **Contribution:** Systematic analysis of parallelization in deep RL across algorithms.

[73] IMPALA: Scalable Distributed Deep-RL with Importance Weighted Actor-Learner Architectures, Espeholt et al, 2018. **Algorithm: IMPALA.**

[74] Distributed Prioritized Experience Replay, Horgan et al, 2018. **Algorithm: Ape-X.**

[75] Recurrent Experience Replay in Distributed Reinforcement Learning, Anonymous, 2018. **Algorithm: R2D2.**

[76] RLlib: Abstractions for Distributed Reinforcement Learning, Liang et al, 2017. **Contribution:** A scalable library of RL algorithm implementations. Documentation link.

9. RL in the Real World

[77] Benchmarking Reinforcement Learning Algorithms on Real-World Robots, Mahmood et al, 2018.

[78] Learning Dexterous In-Hand Manipulation, OpenAI, 2018.

[79] QT-Opt: Scalable Deep Reinforcement Learning for Vision-Based Robotic Manipulation, Kalashnikov et al, 2018. **Algorithm: QT-Opt.**

[80] Horizon: Facebook's Open Source Applied Reinforcement Learning Platform, Gauci et al, 2018.

10. Safety

[81] Concrete Problems in AI Safety, Amodei et al, 2016. **Contribution:** establishes a taxonomy of safety problems, serving as an important jumping-off point for future research. We need to solve these!

[82] Deep Reinforcement Learning From Human Preferences, Christiano et al, 2017. **Algorithm: LFP.**

[83] Constrained Policy Optimization, Achiam et al, 2017. **Algorithm: CPO.**

[84] Safe Exploration in Continuous Action Spaces, Dalal et al, 2018. **Algorithm: DDPG+Safety Layer.**

[85] Trial without Error: Towards Safe Reinforcement Learning via Human Intervention, Saunders et al, 2017. **Algorithm: HIRL.**

[86] Leave No Trace: Learning to Reset for Safe and Autonomous Reinforcement Learning, Eysenbach et al, 2017. **Algorithm: Leave No Trace.**

11. Imitation Learning and Inverse Reinforcement Learning

[87] Modeling Purposeful Adaptive Behavior with the Principle of Maximum Causal Entropy, Ziebart 2010. **Contributions:** Crisp formulation of maximum entropy IRL.

[88] Guided Cost Learning: Deep Inverse Optimal Control via Policy Optimization, Finn et al, 2016. **Algorithm: GCL.**

[89] Generative Adversarial Imitation Learning, Ho and Ermon, 2016. **Algorithm: GAIL.**

[90] DeepMimic: Example-Guided Deep Reinforcement Learning of Physics-Based Character Skills, Peng et al, 2018. **Algorithm: DeepMimic.**

[91] Variational Discriminator Bottleneck: Improving Imitation Learning, Inverse RL, and GANs by Constraining Information Flow, Peng et al, 2018. **Algorithm: VAIL.**

[92] One-Shot High-Fidelity Imitation: Training Large-Scale Deep Nets with RL, Le Paine et al, 2018. **Algorithm: MetaMimic.**

12. Reproducibility, Analysis, and Critique

[93] Benchmarking Deep Reinforcement Learning for Continuous Control, Duan et al, 2016. **Contribution: rllab.**

[94] Reproducibility of Benchmarked Deep Reinforcement Learning Tasks for Continuous Control, Islam et al, 2017.

[95] Deep Reinforcement Learning that Matters, Henderson et al, 2017.

[96] Where Did My Optimum Go?: An Empirical Analysis of Gradient Descent Optimization in Policy Gradient Methods, Henderson et al, 2018.

[97] Are Deep Policy Gradient Algorithms Truly Policy Gradient Algorithms?, Ilyas et al, 2018.

[98] Simple Random Search Provides a Competitive Approach to Reinforcement Learning, Mania et al, 2018.

13. Bonus: Classic Papers in RL Theory or Review

[99] Policy Gradient Methods for Reinforcement Learning with Function Approximation, Sutton et al, 2000. **Contributions:** Established policy gradient theorem and showed convergence of policy gradient algorithm for arbitrary policy classes.

[100] An Analysis of Temporal-Difference Learning with Function Approximation, Tsitsiklis and Van Roy, 1997. **Contributions:** Variety of convergence results and counter-examples for value-learning methods in RL.

[101] Reinforcement Learning of Motor Skills with Policy Gradients, Peters and Schaal, 2008. **Contributions:** Thorough review of policy gradient methods at the time, many of which are still serviceable descriptions of deep RL methods.

[102] Approximately Optimal Approximate Reinforcement Learning, Kakade and Langford, 2002. **Contributions:** Early roots for monotonic improvement theory, later leading to theoretical justification for TRPO and other algorithms.

[103] A Natural Policy Gradient, Kakade, 2002. **Contributions:** Brought natural gradients into RL, later leading to TRPO, ACKTR, and several other methods in deep RL.

[104] Algorithms for Reinforcement Learning, Szepesvari, 2009. **Contributions:** Unbeatable reference on RL before deep RL, containing foundations and theoretical background.