Notes

[Lopes et al., 2012] drive exploration of RL-algorithm wrt progress; use empirical estimate of the "Learning Progress"

Time Dependent MDPs

[Sutton et al., 1999] SMDP
[Boyan and Littman, 2000] TiMDP
[Younes and Simmons, 2004] GSMDP
[Rachelson et al., 2008] XMDP
[Rachelson et al., 2009] TiMDP on UAV
[Bastos et al.,] TiMDP rwd depending on time

UAV

[Baker et al., 2016] Factored MCTS for multi-agent exploration; rescue problem (explore the environment and find the people); toy-domain

[de Marina et al., 2016] Tracking algorithm for smooth dynamic (time varying) trajectories; output = bank angle; real-world system experiments

[Reddy et al., 2016] Learning To Soar; SARSA Learning within a cube including a turbulent windfield with thermal phenomenon

[Fechner et al., 2015] Dynamic model of a pumping kite power system

[Lawrance, 2011] TODO - Autonomous soaring flight for unmanned aerial vehicles; PhD Thesis; use GP to model belief over windfield

MCTS/Bandit problem

[Kocsis and Szepesvári, 2006] UCT

[Browne et al., 2012] General survey of MCTS

[Couëtoux et al., 2011] PW and DPW for continuous MCTS; use finite discrete set of actions and s'

[Bubeck et al., 2009] HOO (Hierarchical Optimistic Optimization): build trees of covering $(P_{h,i})_{1<=i<=2^h}$; each h corresponds to a level of dividing A the set of actions giving a binary tree; UCT-like approach to explore the binary tree composed with subsets of A; almost a dichotomy approach

[Mansley et al., 2011] HOOT (HOO for Trees)

[Weinstein and Littman, 2012] HOLOP (Hierarchical Open-Loop Optimistic Planning) application of HOO to MDPs for sequential planning

[Auger et al., 2013] PUCT (Polynomial UCT); no knowledge of A (black-box); tree of decision $z : (s)$ and random $w : (s, a)$ nodes; use DPW + PUCT i.e. polynomial exploration with the following UCT score:

$$a_{next} = argmax_a \left\{ \widehat{V}(z, a) + \sqrt{\frac{n(z)^{e(depth)}}{n(z, a)}} \right\}$$

[Silver et al., 2008] permanent and transient memory; SARSA + dyna architecture

[Deleva, 2015] TD learning within MCTS (PhD thesis)

[Gelly and Silver, 2011] MCTS + RAVE (Rapid Action Value Estimation); 2 features: RAVE + heuristic function used to initialize value of nodes; RAVE = bias of the value estimate of node z based on the subtree $\tau(z)$; the bias is the all-move-as-first Q value: mean of every same move's Q value played in the subtree; heuristic function = TD learning with self-play

Actor-Critic

[Lagoudakis and Parr, 2003] LSPI
[Xu et al., 2007] Kernelized LSPI
[Busoniu et al., 2010] include online LSPI

Caption

- PW: Progressive Widening
- DPW: Double Progressive Widening

References

[Auger et al., 2013] Auger, D., Couetoux, A., and Teytaud, O. (2013). Continuous upper confidence trees with polynomial exploration–consistency. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 194–209. Springer.

[Baker et al., 2016] Baker, C., Ramchurn, G., Teacy, L., and Jennings, N. (2016). Factored monte-carlo tree search for coordinating uavs in disaster response. ICAPS.

[Bastos et al.,] Bastos, G. S., Ramos, F. T., de Souza, L. E., and Ribeiro, C. H. Time-dependent utility decision making using the timdp model.

[Boyan and Littman, 2000] Boyan, J. A. and Littman, M. L. (2000). Exact solutions to time-dependent mdps. In *NIPS*, pages 1026–1032.

[Browne et al., 2012] Browne, C. B., Powley, E., Whitehouse, D., Lucas, S. M., Cowling, P. I., Rohlfshagen, P., Tavener, S., Perez, D., Samothrakis, S., and Colton, S. (2012). A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in Games*, 4(1):1–43.

[Bubeck et al., 2009] Bubeck, S., Stoltz, G., Szepesvári, C., and Munos, R. (2009). Online optimization in x-armed bandits. In *Advances in Neural Information Processing Systems*, pages 201–208.

[Busoniu et al., 2010] Busoniu, L., Babuska, R., De Schutter, B., and Ernst, D. (2010). *Reinforcement learning and dynamic programming using function approximators*, volume 39. CRC press.

[Couëtoux et al., 2011] Couëtoux, A., Hoock, J.-B., Sokolovska, N., Teytaud, O., and Bonnard, N. (2011). Continuous upper confidence trees. In *International Conference on Learning and Intelligent Optimization*, pages 433–445. Springer.

[de Marina et al., 2016] de Marina, H. G., Kapitanyuk, Y. A., Bronz, M., Hattenberger, G., and Cao, M. (2016). Guidance algorithm for smooth trajectory tracking of a fixed wing uav flying in wind flows. *arXiv preprint arXiv:1610.02797*.

[Deleva, 2015] Deleva, A. (2015). *TD Learning in Monte Carlo Tree Search: Masters Thesis*. PhD thesis, A. Deleva.

[Fechner et al., 2015] Fechner, U., van der Vlugt, R., Schreuder, E., and Schmehl, R. (2015). Dynamic model of a pumping kite power system. *Renewable Energy*, 83:705–716.

[Gelly and Silver, 2011] Gelly, S. and Silver, D. (2011). Monte-carlo tree search and rapid action value estimation in computer go. *Artificial Intelligence*, 175(11):1856–1875.

[Kocsis and Szepesvári, 2006] Kocsis, L. and Szepesvári, C. (2006). Bandit based monte-carlo planning. In *European conference on machine learning*, pages 282–293. Springer.

[Lagoudakis and Parr, 2003] Lagoudakis, M. G. and Parr, R. (2003). Least-squares policy iteration. *Journal of Machine Learning Research*, 4(Dec):1107–1149.

[Lawrance, 2011] Lawrance, N. R. (2011). *Autonomous soaring flight for unmanned aerial vehicles*. University of Sydney.

[Lopes et al., 2012] Lopes, M., Lang, T., Toussaint, M., and Oudeyer, P.-Y. (2012). Exploration in model-based reinforcement learning by empirically estimating learning progress. In *Advances in Neural Information Processing Systems*, pages 206–214.

[Mansley et al., 2011] Mansley, C. R., Weinstein, A., and Littman, M. L. (2011). Sample-based planning for continuous action markov decision processes. In *ICAPS*.

[Rachelson et al., 2009] Rachelson, E., Fabiani, P., and Garcia, F. (2009). Adapting an mdp planner to time-dependency: case study on a uav coordination problem. In *4th Workshop on Planning and Plan Execution for Real-World Systems: Principles and Practices for Planning in Execution, Thessaloniki, Greece*.

[Rachelson et al., 2008] Rachelson, E., Garcia, F., and Fabiani, P. (2008). Extending the bellman equation for mdp to continuous actions and continuous time in the discounted case. In *10th Int. Symp. on AI and Math.*

[Reddy et al., 2016] Reddy, G., Celani, A., Sejnowski, T. J., and Vergassola, M. (2016). Learning to soar in turbulent environments. *Proceedings of the National Academy of Sciences*, page 201606075.

[Silver et al., 2008] Silver, D., Sutton, R. S., and Müller, M. (2008). Sample-based learning and search with permanent and transient memories. In *Proceedings of the 25th international conference on Machine learning*, pages 968–975. ACM.

[Sutton et al., 1999] Sutton, R. S., Precup, D., and Singh, S. (1999). Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211.

[Weinstein and Littman, 2012] Weinstein, A. and Littman, M. L. (2012). Bandit-based planning and learning in continuous-action markov decision processes. In *ICAPS*.

[Xu et al., 2007] Xu, X., Hu, D., and Lu, X. (2007). Kernel-based least squares policy iteration for reinforcement learning. *IEEE Transactions on Neural Networks*, 18(4):973–992.

[Younes and Simmons, 2004] Younes, H. L. and Simmons, R. G. (2004). Solving generalized semi-markov decision processes using continuous phase-type distributions. In *AAAI*, volume 4, page 742.