强化学习和蒙特卡洛树搜索 (MCTS)

1 强化学习 (reinforcement learning)

1.1 强化学习能做什么

图 1: AlphaGo vs 柯洁

图 2: OpenAI Dota2

1.2 强化学习和无监督学习、有监督学习的关系和区别

- 有监督学习：学习从特征到 label 的映射

- 无监督学习：从无标记样本中发现样本隐藏的结构
- 强化学习：最大化 reward

总的来说，RL 与其他机器学习算法不同的地方在于：其中没有监督者，只有一个 reward 信号；reward 可能是延迟的；时间在 RL 中具有重要的意义；agent 的行为会影响之后一系列的 data。

1.3 强化学习的关键词和术语

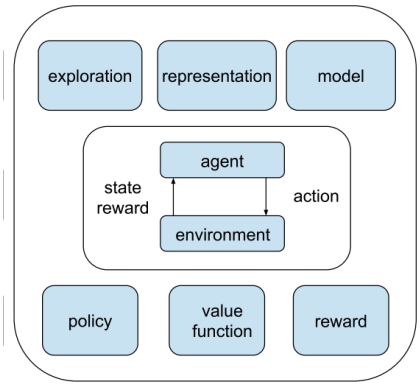

图 3: 强化学习框架

- states and obervation
- action space
- policy
- trajectory
- different formulations of return
- the RL optimization problem
- value function

2 蒙特卡洛树搜索

2.1 蒙特卡洛方法

蒙特卡洛方法 (Monte Carlo Methods) 是强化学习中基于无模型的训练方法。与动态规划 (Dynamic Programming) 不同，该方法并没有明确的模型 (即 transition-state probability)，也就是说我们并不知道各个状态之间转换的概率，可以把它看作是环境 (environment) 模型。

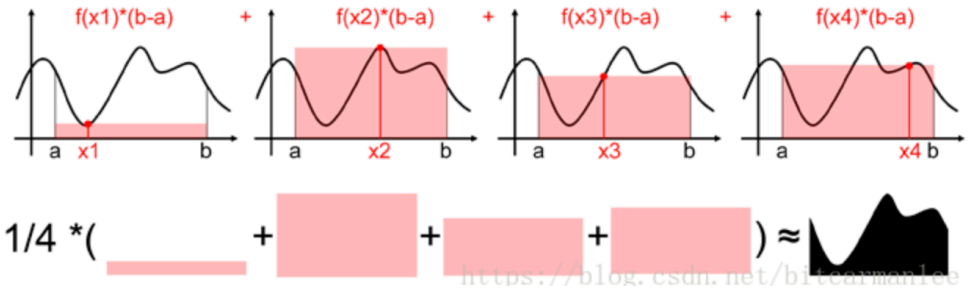

图 4: 蒙特卡洛方法求定积分

2.2 蒙特卡洛树搜索

蒙特卡洛树搜索（Monte Carlo Tree Search），是一类树搜索算法的统称，可以较为有效地解决一些探索空间巨大的问题。

要求的条件是 zero-sum、fully information、determinism、sequential、discrete，也即是说这个场景必须是能分出输赢（不能同时赢）、游戏的信息是完全公开的（不像打牌可以隐藏自己的手牌）、确定性的（每一个操作结果没有随机因素）、顺序的（操作都是按顺序执行的）、离散的（动作空间是有限的集合）

3 深度强化学习和蒙特卡洛搜索树的结合

神经网络的 loss function:

$$(p, v) = f_\theta(s) and l = (z - v)^2 - \pi^T log(p) + c||\theta||^2$$

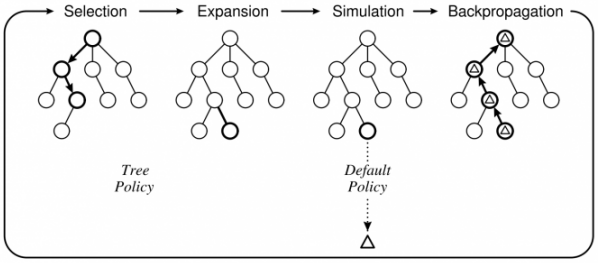

图 5: 蒙特卡洛树搜索的一次 playout/rollout

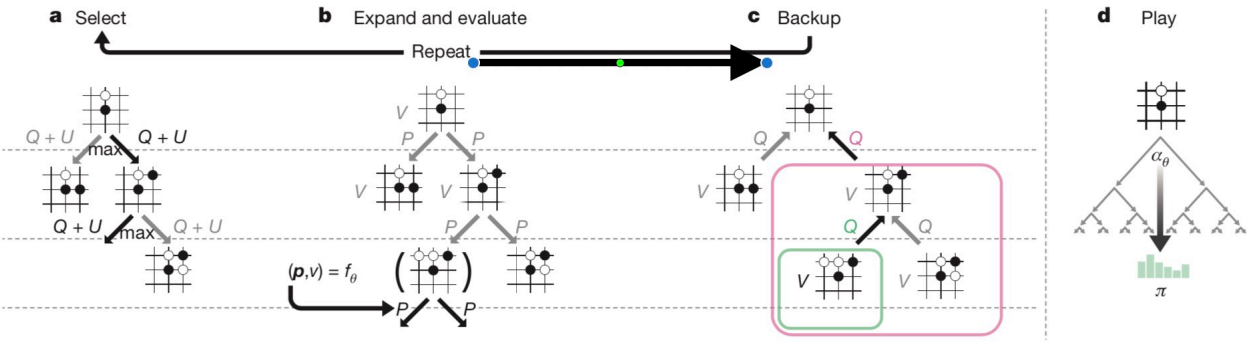

图 6: AlphaGo Zero 中的蒙特卡洛搜索树

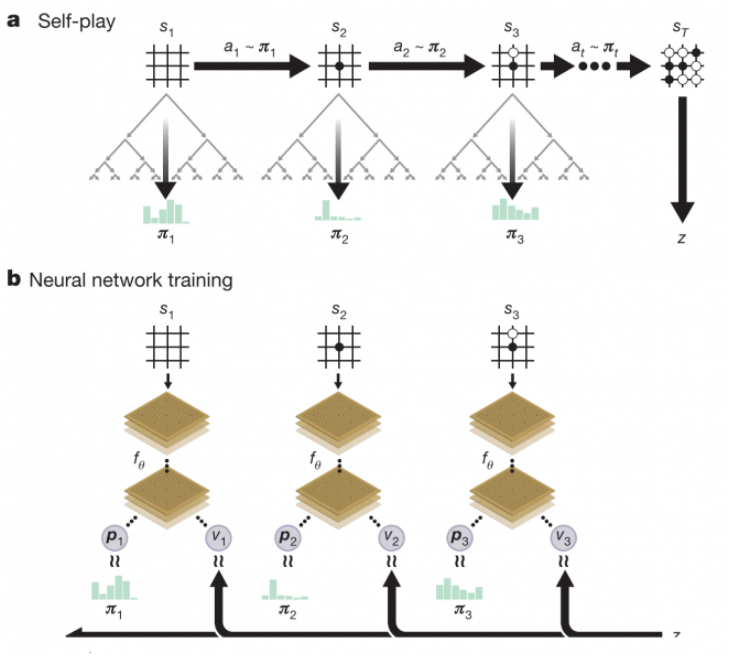

图 7: AlphaGo Zero 的神经网络训练过程

4 参考文献

1. Li, Y. (2018). Deep Reinforcement Learning, 1–150

2. Browne, C. B., Powley, E., Whitehouse, D., Lucas, S. M., Cowling, P. I., Rohlfshagen, P., ⋯Colton, S. (2012). A survey of Monte Carlo tree search methods. IEEE Transactions on Computational Intelligence and AI in Games, 4(1), 1–43. https://doi.org/10.1109/TCIAIG.2012.2186810

3. Lanctot, M., Hassabis, D., Graepel, T., Panneershelvam, V., Lillicrap, T., Nham, J., ⋯Dieleman, S. (2016). Mastering the game of Go with deep neural networks and tree search. Nature, 529(7587), 484 – 489. https://doi.org/10.1038/nature16961

4. Hubert, T., Schrittwieser, J., Baker, L., Hui, F., Hassabis, D., Antonoglou, I., ⋯Guez, A. (2017). Mastering the game of Go without human knowledge. Nature, 550(7676), 354–359. https://doi.org/10.1038/nature24270

5. Kumaran, D., Hassabis, D., Graepel, T., Lai, M., Silver, D., Lanctot, M., ⋯Guez, A. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. Science, 362(6419), 1140 – 1144. https://doi.org/10.1126/science.aar6404

6. Tian, Y., Ma, J., Gong, Q., Sengupta, S., Chen, Z., Pinkerton, J., & Zitnick, C. L. (2019). ELF OpenGo: An Analysis and Open Reimplementation of AlphaZero. Retrieved from https://arxiv.org/abs/1902.04522