Bayesian Mixture Modeling and Inference based Thompson Sampling in Monte-Carlo Tree Search

Aijun Bai†, Feng Wu‡, and Xiaoping Chen†
University of Science and Technology of China†, and University of Southampton‡

Background

Monte-Carlo tree search (MCTS) finds near-optimal policies in domains of online planning for *Markov decision processes* (MDPs) by combining tree search methods with sampling techniques. The key idea is to iteratively evaluate each state in a best-first search tree by the mean outcome of simulation samples.

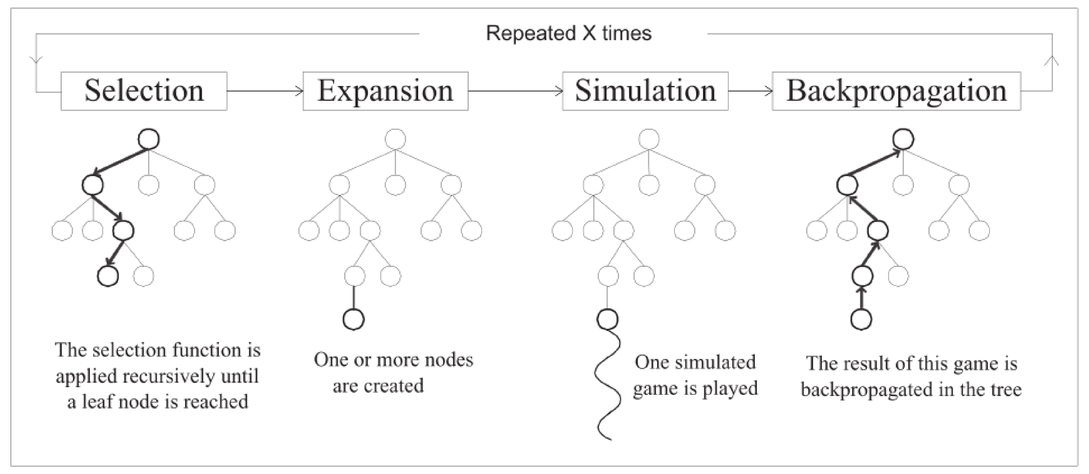

Figure 1: Monte-Carlo Tree Search [Chaslot2008]

Motivation

When applying MCTS, one of the fundamental challenges is the so-called *exploration versus exploitation* dilemma. Thompson sampling selects actions stochastically, based on the probabilities of being optimal. In this paper, we borrow the idea of Thompson sampling and propose the Dirichlet-NormalGamma MCTS (DNG-MCTS) algorithm.

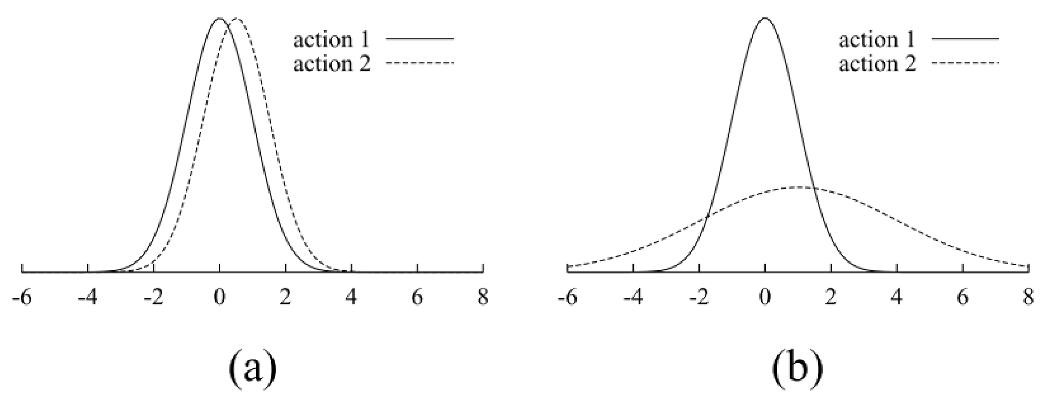

Figure 2: Thompson sampling based action selection [Dearden1998]

Assumptions

For a given MDP policy π, let random variables:

1. $X_{s,\pi}$ denotes the accumulated reward of following policy π starting from state s,
2. $X_{s,a,\pi}$ denotes the accumulated reward of first performing action a in state s and then following policy π thereafter.

According to the central limit theorem on Markov chains, our assumptions are:

1. $X_{s,\pi}$ is sampled from a Normal distribution,
2. $X_{s,a,\pi}$ can be modeled as a mixture of Normal distributions.

Algorithm

DNG-MCTS algorithm extends MCTS:

- Model the unknown distribution of $X_{s,a,\pi}$ as a mixture of Normal distributions,
- Choose the conjugate prior in the form of a combination of Dirichlet and Normal-Gamma distributions,
- Compute the posterior distribution after each accumulated reward is observed by simulation in the search tree,
- Use Thompson sampling to guide the selection of actions at each decision node.

Conclusion and Future Work

DNG-MCTS algorithm:

- Monte-Carlo tree search framework
- Bayesian mixture modeling and inference
- Thompson sampling based action selection
- Competitive results comparing to UCT

In the future, we plan to extend our basic assumptions to using more realistic distributions and test our algorithm on real-world applications.

Experiments

We have tested DNG-MCTS and compared the results with UCT in three benchmark domains, namely *Canadian traveler problem*, *racetrack* and *sailing*. In each benchmark problem, we:

1. Ran the algorithms for a number of iterations from the current state,
2. Applied the best action based on the resulted action-values,
3. Repeated the loop until terminating conditions,
4. Reported the total discounted cost.

DNG-MCTS produced competitive results in *CTP*, and converged faster in *racetrack* and *sailing* with respect to sample complexity.

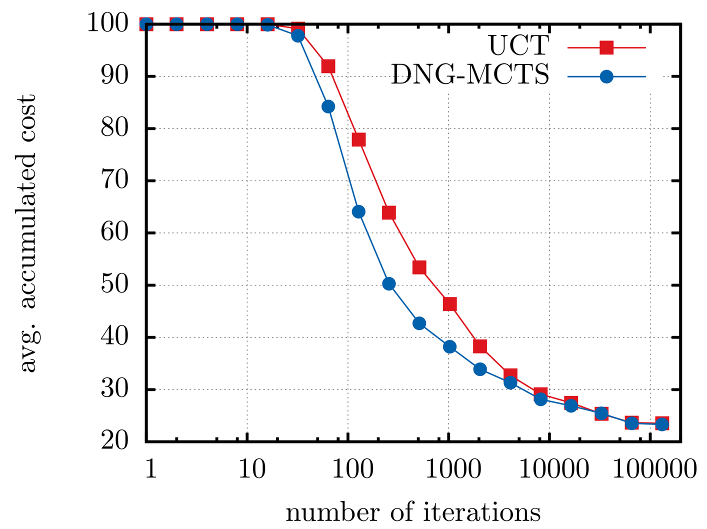

Figure 3: Racetrack-barto-big with random policy

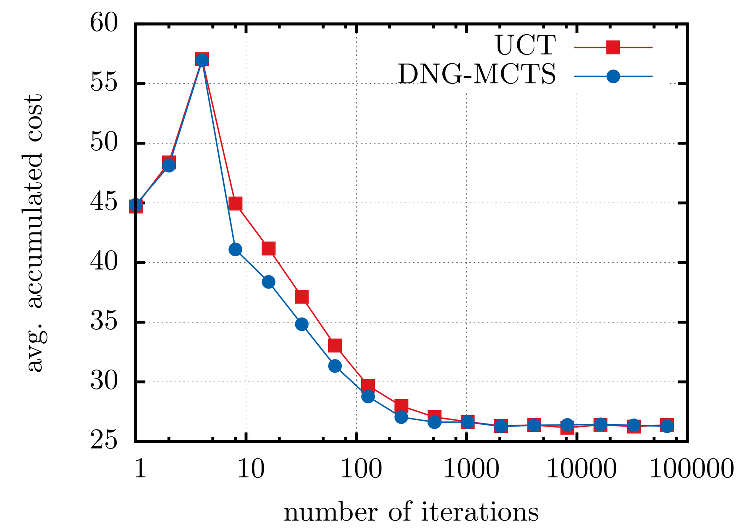

Figure 4: Sailing-100 × 100 with random policy

Table 1: CTP problems with 20 nodes. The second column indicates the belief size of the transformed MDP for each problem instance. UCTB and UCTO are the two domain-specific UCT implementations. DNG-MCTS and UCT run for 10,000 iterations. Boldface fonts are best in whole table; gray cells show best among domain-independent implementations for each group.

		domain-specific UCT		random rollout policy		optimistic rollout policy	
prob.	belief	UCTB	UCTO	UCT	DNG	UCT	DNG
20-1	20×3^{49}	210.7±7	**169.0±6**	216.4±3	223.9±4	180.7±3	177.1±3
20-2	20×3^{49}	176.4±4	**148.9±3**	178.5±2	178.1±2	160.8±2	155.2±2
20-3	20×3^{51}	150.7±7	**132.5±6**	169.7±4	159.5±4	144.3±3	140.1±3
20-4	20×3^{49}	264.8±9	**235.2±7**	264.1±4	266.8±4	238.3±3	242.7±4
20-5	20×3^{52}	123.2±7	**111.3±5**	139.8±4	133.4±4	123.9±3	122.1±3
20-6	20×3^{49}	165.4±6	**133.1±3**	178.0±3	169.8±3	167.8±2	141.9±2
20-7	20×3^{50}	191.6±6	**148.2±4**	211.8±3	214.9±4	174.1±2	166.1±3
20-8	20×3^{51}	160.1±7	**134.5±5**	218.5±4	202.3±4	152.3±3	151.4±3
20-9	20×3^{50}	235.2±6	**173.9±4**	251.9±3	246.0±3	185.2±2	180.4±2
20-10	20×3^{49}	180.8±7	**167.0±5**	185.7±3	188.9±4	178.5±3	170.5±3
total		1858.9	**1553.6**	2014.4	1983.68	1705.9	1647.4

Contact Information

Web http://home.ustc.edu.cn/~baj

Email baj@mail.ustc.edu.cn