Monte-Carlo tree search with uncertainty propagation via optimal transport

1st Shri Prathaa
Electrical engineering
IIT Madras

1st Clifford.F
IDDD Data Science
IIT Madras

***Abstract*—In many situations, like solving complex games or reinforcement learning tasks, Monte Carlo Tree Search (MCTS) is useful, but in highly stochastic and partially observable Markov decision processes (MDP), propagation of simple mean to the root node isn't enough due to high variability. Instead, we model value and action-value nodes as Gaussian distributions. We introduce a new backup operator that computes values using the Wasserstein barycenter, which propagates uncertainty across the tree. This probabilistic approach is enhanced by two sampling strategies: Optimistic selection and Thompson sampling, leading to a Wasserstein MCTS algorithm.**

I. Introduction

Reinforcement Learning (RL) is of two major types: model-based (typically using a Markov Decision Process such as Monte Carlo tree search) or model-free (such as using the hidden layers of a deep neural network to approximate the value function, e.g., temporal difference learning). However, in highly stochastic environments, possibly with partial observability, value estimation through deep learning alone isn't helpful. In normal MCTS, high uncertainty can cause suboptimal action selection and poor performance. Therefore, we model action values as Gaussian distributions and use Wasserstein barycenters with α-divergence as the distance measure. This operator adopts a power mean backup that addresses the value overestimation issue.

II. Related Work

The use of L_2-Wasserstein barycenters in combination with Euclidean distance to propagate the uncertainty of the action-value estimate is found in Metelli et al. (2019).

Tesauro et al. (2012) propose representing the value function at each node as a Gaussian distribution and using the standard deviation of the Gaussian for better estimation of the exploration constant.

Bai et al. (2013) model the value function using a Dirichlet-NormalGamma distribution(DNG) coupled with Thompson sampling exploration for highly stochastic environments, and Bai et al. (2014) extend this method to partially observable problems.

Generalized mean backup operators are extensively studied in Dam et al. (2019) and Coulom (2007) to tackle the under-estimation/overestimation problems of average and maximum backup operators in UCT.

The two strategies—optimism in the face of uncertainty (Auer et al., 2002a) and Thompson sampling (Thompson, 1933)—are our sampling methods.

III. Problem Setup

The paper focuses on solving stochastic and partially observable MDPs. An MDP is represented by a tuple $M = (S, A, R, P, \gamma)$, where S is the state space, A is the action space, R is the reward function, P is the state transition function, and γ is the discount factor. The goal is to find the optimal policy π that maximizes the expected cumulative reward calculated using the Bellman equation.

In MCTS, the tree's nodes correspond to states, and the edges represent actions. Standard MCTS uses a backup operator to propagate values from child nodes to parent nodes. However, in stochastic environments, this approach can lead to overestimation or underestimation of the value function. To address this, the proposed method models value nodes as Gaussian distributions, where uncertainty is propagated using the L_1-Wasserstein barycenter.

IV. Algorithms

MCTS has 4 steps: Selection, Expansion, Simulation, and Backup. The tree-policy used to select the action should balance exploration and exploitation. The reached node is expanded and a rollout is performed. Then, the value is propagated to its roots.

A. Upper Confidence bound for trees

The upper confidence bounds for trees (UCT) algorithm is an extension of the UCB1 multi-armed bandit algorithm. UCB1 chooses the arm (action a) using:

$$a = \arg \max_{i \in \{1,\dots,K\}} \left(\overline{X}_{i,T_i(n-1)} + C\sqrt{\frac{\log n}{T_i(n-1)}} \right),$$

where $T_i(n) = \sum_{t=1}^{n} \mathbb{I}\{t = i\}$ is the number of times arm i is played up to time n, $\overline{X}_{i,T_i(n-1)}$ denotes the average reward of arm i up to $n-1$, and $C = \sqrt{2}$ is an exploration constant. The value is backed up recursively from the leaf node to the root as:

$$\overline{X}_n = \sum_{i=1}^{K} \frac{T_i(n)}{n} \overline{X}_{i,T_i(n)}.$$

B. Wasserstein Barycenter

Given a complete separable metric space (X, d), the **Lq-Wasserstein distance** between two probability measures μ and ν is defined as:

$$W_q(\mu,\nu) = \left(\inf_{\rho\in\Gamma(\mu,\nu)} \mathbb{E}_{X,Y\sim\rho}[d(X,Y)^q]\right)^{1/q}$$

where $\Gamma(\mu,\nu)$ represents the set of all couplings of μ and ν. The **Lq-Wasserstein barycenter** is the probability measure $\bar{\nu}$ that minimizes the (weighted) Wasserstein distance between a set of measures $\{\nu_i\}$:

$$\bar{\nu} = \arg\inf_{\nu\in N} \sum_{i=1}^{n} w_i W_q(\nu,\nu_i)^q$$

C. α-Divergence

The α**-divergence** is a distance measure between two points X and Y on a manifold M with coordinates $\xi_X^{(i)}$ and $\xi_Y^{(i)}$

$$D_f^\alpha(X \parallel Y) = \sum_i \xi_Y^{(i)} f_\alpha\left(\frac{\xi_X^{(i)}}{\xi_Y^{(i)}}\right)$$

where the α**-function** is given by:

$$f_\alpha(x) = \frac{(x^\alpha - 1) - \alpha(x-1)}{\alpha(\alpha-1)}$$

The **L1-Wasserstein barycenter** with α-divergence as distance measure is used to propagate uncertainty of value estimates across the tree.

D. V-Posterior

The **V-posterior** represents the value of a node $V(s)$ as the **L1-Wasserstein barycenter** of the child **Q-posteriors** $Q(s,a)$, given a policy $\bar{\pi}$:

$$V(s) = \arg\inf_V \mathbb{E}_{a\sim\bar{\pi}(.|s)} W_1(V, Q(s,a))$$

Proposition 1
If $V(s)$ and $Q(s,a)$ are modeled as **Gaussian distributions** $N(m(s),\sigma^2(s))$, then the V-posterior is computed as:

$$m(s) = \left(\mathbb{E}_{a\sim\bar{\pi}}[m(s,a)^p]\right)^{1/p}, \quad \sigma(s) = \left(\mathbb{E}_{a\sim\bar{\pi}}[\sigma(s,a)^p]\right)^{1/p}$$

This gives the mean and standard deviation of $V(s)$ as the **power mean** of the Q-posteriors.

Proposition 2
If the nodes are modeled as a **particle model** with M particles, each particle $x_i(s)$ of $V(s)$ is derived as:

$$x_i(s) = \left(\mathbb{E}_{a\sim\bar{\pi}}[x_i(s,a)^p]\right)^{1/p}$$

This is a particle-based generalization of Proposition 1 using the same power mean concept.

E. Backup Operator

In the W-MCTS algorithm, the following mean and standard deviation value backup operator of the V-node is proposed considering how frequently each action (Q-node) has been taken relative to the total number of visits

$$\overline{V}_m(s,N(s)) \leftarrow \left(\sum_a \frac{n(s,a)}{N(s)} \overline{Q}_m(s,a,n(s,a))^p\right)^{\frac{1}{p}}.$$

$$\overline{V}_{\text{std}}(s,N(s)) \leftarrow \left(\sum_a \frac{n(s,a)}{N(s)} \overline{Q}_{\text{std}}(s,a,n(s,a))^p\right)^{\frac{1}{p}}.$$

where $\overline{V}_m(s,N(s))$ and $\overline{Q}_m(s,a,n(s,a))$ denote the empirical mean value for V- and Q-nodes after $N(s)$, and $n(s,a)$ visitations and likewise for standard deviation.

For the backup operators of Q_m and Q_{std}, Bellman operator is used and visitation ratio for policy selection is employed to get the backup operators

$$\overline{Q}_m(s,a,n(s,a)) \leftarrow \frac{\sum r(s,a) + \gamma \sum_{s'} N(s')\overline{V}_m(s',N(s'))}{n(s,a)},$$

$$\overline{Q}_{\text{std}}(s,a,n(s,a)) \leftarrow \frac{\gamma \sum_{s'} N(s')\overline{V}_{\text{std}}(s',N(s'))}{n(s,a)}.$$

F. Action Selection

1) Optimistic Selection: Action is selected to maximize the statistical upper confidence limits of lower level Q-nodes according to UCT, as

$$a = \arg\max_{a_i\in\{1,\ldots,K\}} \left(m(s,a_i) + C\sigma_i(s,a_i)\sqrt{\log N(s)}\right).$$

with $i = 1,2,\ldots,K$ is the action index, $N(s)$ is the visitation count of the current V node at state s and C is an exploration constant. The exploration term $1/\sqrt{n_i(s,a_i)}$ in UCT is replaced by the standard deviation of the $Q(s,a_i)$ node to derive Wasserstein Monte-Carlo tree search using optimistic selection (*W-MCTS-OS*). This comes from the central limit theorem as $\sigma_i(s,a_i)^2 \sim 1/n_i(s,a_i)$.

2) Thompson Sampling: The action is selected based on the posteriors of the action-value distribution as :

$$a = \arg\max_{a_i\in\{1,\ldots,K\}} \{\theta_i \sim \mathcal{N}(m(s,a_i),\sigma^2(s,a_i))\}.$$

Here, $\mathcal{N}(m(s,a_i),\sigma^2(s,a_i))$ is a Gaussian distribution over following Q-nodes with mean value $m(s,a_i)$ and standard deviation value $\sigma(s,a_i)$, for each action a_i at state s. This method is denoted as Wasserstein Monte-Carlo tree search Thompson Sampling (*W-MCTS-TS*).

V. Main Results

A. Theoretical Analysis

Consider $K \geq 1$ arms or actions of interest with mean value $\mu_k, k \in [K]$. Let $X_{k,t}$ denote the random reward obtained by playing arm $k \in [K]$ at the time step t, $\overline{X}_{k,n} = \frac{1}{n}\sum_{t=1}^{n} X_{k,t}$ is the average reward and $\mu_{k,n} = \mathbb{E}[\overline{X}_{k,n}]$. $T_k(n)$ denotes the number of visits of arm k.

Assumption 1. We assume that the reward sequence, $\{X_{k,t} : t \geq 1\}$, is a non-stationary process satisfying the assumption:

1. (Gaussian) Each arm k is a Gaussian $\mathcal{N}(\mu_k, V_k/T_k(n))$, $V_i > 0$.

2. (Convergence) The expectation $\mu_{k,n}$ converges to a value μ_k

$$\mu_k = \lim_{n\to\infty} \mathbb{E}[\overline{X}_{kn}].$$

Consider Thompson Sampling strategy applied to a non-stationary problem where the pay-off sequence satisfies Assumption 1. Define $V = \max_{k\in[K]}\{V_k\}$,$\Delta_k = \mu^* - \mu_k$ and $\Delta = \max_{k\in[|K|]}\{\Delta_k\}$. Fix $\epsilon \geq 0$. If k is the index of a sub-optimal arm, then each sub-optimal arm k is played in expectation at most

$$E[T_k(n)] \leq \Theta\left(1 + \frac{V\log(n\Delta_k^2/V)}{\Delta_k^2}\right).$$

B. Convergence in Monte-Calo Tree Search

The probability of not choosing an optimal action at the root node decays polynomially to zero. At the root node, let a_k be the action returned by W-MCTS-TS at timestep n, a_{k^*} is the optimal action. Then for $\epsilon > 0$, $\exists C > 0$, $\alpha > 0$, $N_p > 0$ as constants that for all $n \geq N_p$, we have

$$\Pr(a_k \neq a_{k^*}) \leq Cn^{-\alpha}.$$

Next, Polynomial convergence of the expected estimated mean value function at the root node is shown.

At the root node $s^{(0)}$, $\exists N_0 > 0$, so that the expected payoff satisfies

$$\left|\mathbb{E}\left[V_m(s^{(0)}, n)\right] - Q_m(s^{(0)}, a_{k^*})\right| \leq \Theta\left(\frac{2(|K|-1)\left(1 + \frac{V\log(n\Delta^2/V)}{\Delta^2}\right)}{n}\right).$$

W-MCTS-TS, guarantees a polynomial convergence rate to the original optimal value function at the root node.

VI. Simulations

The results of different variations of W-MCTS are compared with UCT, Power-UCT, DNG, and D2NG. Power-UCT is a variant of UCT based on the power mean operator, and D2NG is an extension of DNG where the posterior distribution is modeled by choosing the conjugate prior in the form of a combination of two Dirichlet and one NormalGamma distribution.

A. Fully observable highly-stochastic problems

FrozenLake, NChain, RiverSwim, SixArm, and Taxi environments were used for comparison. W-MCTS consistently outperforms UCT, Power-UCT, and other methods. In FrozenLake and NChain, Wasserstein sampling (especially with Thompson sampling) shows faster convergence and superior performance. On RiverSwim, W-MCTS with optimistic selection achieves the best results. Only W-MCTS with Thompson sampling could handle the large state space in the complex Taxi environment.

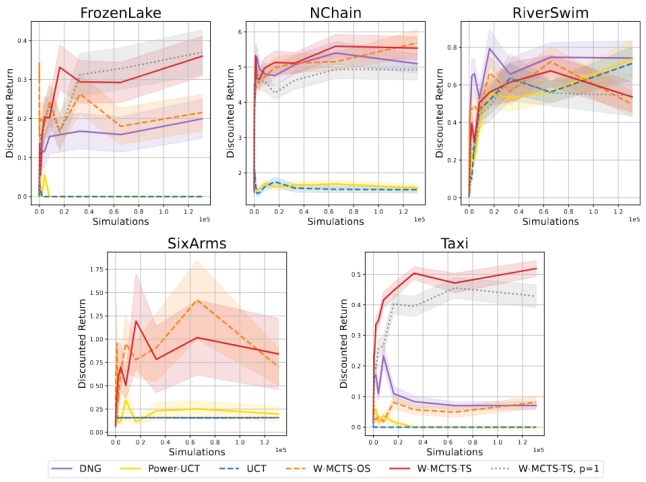

Fig. 1. Performance of W-MCTS compared to DNG, Power-UCT and UCT in different MDP environments. The mean of total discounted reward over 50 evaluation runs is shown by thick lines while the shaded area shows standard error.

B. Partially observable highly-stochastic problems

The W-MCTS method was compared with D2NG and UCT in 2 environments- Rocksample and Pocman. Results in Figure 2 W-MCTS with Thompson sampling consistently outperforms UCT and D2NG across all action settings (16, 20, and 40 actions) in the Rocksample environment. Similar results can be seen in the PocMan environment where W-MCTS with Thompson sampling (p=100) shows superior performance, surpassing UCT and D2NG with a range of sample sizes (4096, 32768, and 65536).

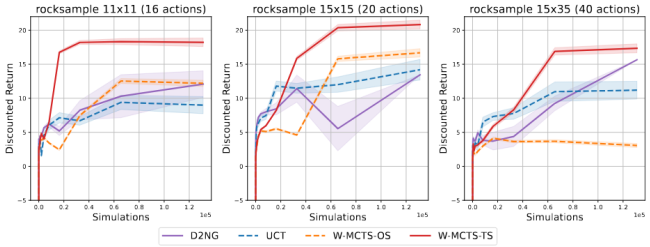

Fig. 2. Performance of W-MCTS compared to D2NG in rocksample. The mean of total discounted reward over 1000 (except for UCT with 100) evaluation runs is shown by thick lines while the shaded area shows standard error.

VII. Conclusion

This paper introduces a novel probabilistic approach to Monte-Carlo tree search (MCTS) for highly stochastic MDPs and Partially Observable MDPs. It models value nodes as Gaussian distributions and adopts a novel backup operator that computes value nodes as Wasserstein barycenters of children action-value nodes. We choose actions using Thompson sampling and optimistic selection. It is also observed that using Thompson sampling to choose the optimal action at the root node shows polynomial convergence.