Maximum Entropy Monte-Carlo Planning
Chenjun Xiao¹², Jincheng Mei¹², Ruitong Huang³, Dale Schuurmans¹², Martin Müller¹²

Entropy Regularized Value Functions

Maximum entropy policy optimization

\[ \max_{\pi} \left\{ \pi \cdot r + \tau H(\pi) \right\}, \]

where \( r \in \mathbb{R}^K \) is the reward vector and \( \tau \geq 0 \) is the temperature.

**Note** The optimal solution is given by *softmax*.

\[ F_\tau(r) = \max \left\{ \pi \cdot r + \tau H(\pi) \right\} = f_\tau(r) \cdot r + \tau H(f_\tau(r)). \]

**Motivation** Use the smoothed softmax value in Monte-Carlo planning.

**Softmax policy**

\[ \pi_\text{sm}(a|s) = \exp \left\{ \frac{(Q_\text{sm}(s,a) - V_\text{sm}(s))/\tau}{R} \right\} \]

**Softmax value functions**

\[ Q_\text{sm}(s,a) = R(s,a) + \frac{\mathbb{E}_{\tau|s,a} [ V_\text{sm}(s')]}{\tau} \quad V_\text{sm}(s) = \tau \log \sum_a \exp \left\{ Q_\text{sm}(s,a)/\tau \right\} \]

Stochastic Softmax Bandit

- Sequential decision making
- \( K \) actions, (unknown) expected rewards \( r \in \mathbb{R}^K \)
- At each round \( t \), play \( A_t \) and receive \( R_t \) (subgaussian)
- Goal: estimate \( F_\tau(r) \)

We use an estimator constructed by empirical means \( V_t = \tau \log U_t \), where \( U_t = 1 \cdot e^{\tau/\epsilon} \). Let \( U^* = 1 \cdot e^{\tau/\epsilon} \) and \( V^* = \tau \log U^* \).

**Theorem** We aim to minimize mean squared error \( E_t = \mathbb{E}[(U^* - U_t)^2] \)

**Theorem (lower bound)** For any algorithm that achieve \( E_t = O(1/t) \), there exists a problem setting such that

\[ \lim_{t \to \infty} t E_t \geq \frac{\sigma^2}{2\epsilon} \left( 1 - e^{\tau/\epsilon} \right)^2. \]

Furthermore, to achieve this lower bound, there must be for any \( a \in A \), \( \lim_{t \to \infty} N_t(a)/t = \pi_{\text{sm}}(a) \).

Optimal Sequential Sampling

**Empirical Exponential Weight (E2W)**

\[ s_t(a) = (1 - \lambda_t) f_\tau(f_\tau(a)) + \lambda_t |A| \]

where \( \lambda_t = \epsilon |A| / \log(t + 1) \).

**Theorem** E2W is asymptotically optimal, i.e.

\[ \lim_{t \to \infty} t E_t = \frac{\sigma^2}{2\epsilon} \left( 1 - e^{\tau/\epsilon} \right)^2. \]

Maximum Entropy MCTS

**Main Idea**

- Use E2W as in-tree policy
- Use softmax value functions as state value

**Softmax value backpropagation** let \( \{s_1, a_0, \ldots, s_T\} \) be the state action trajectory in a simulation, \( R \) be the return of an evaluation on \( s_T \).

\[ Q_{\text{sm}}(s_t, a_t) = \begin{cases} r(s_t, a_t) + R & t = T - 1 \\ r(s_t, a_t) + F_\tau(Q_{\text{sm}}(s_{t+1})) & t < T - 1 \end{cases} \]

**Theorem** For any state \( s \) and action \( a \), if the algorithm explores actions according to \( \pi_{\text{sm}} \), i.e. \( N^*(s,a) = \pi_{\text{sm}}(a|s) \cdot N(s) \), then for \( \epsilon \in [0,1] \),

\[ \mathbb{P} \left\{ |V_{\text{sm}}(s) - V_{\text{sm}}(s')| \geq \epsilon \right\} \leq C \exp \left( -N(s) \epsilon^2 / C \sigma^2 \right) \]

with some constant \( C \).

**Theorem** Let \( a_t \) be the action returned by MENTS at iteration \( t \). Then for large enough \( t \),

\[ \mathbb{P} \left\{ a_t \neq a^* \right\} \leq C \exp \left( -t / (\log t)^3 \right) \]

with some constant \( C \).

Experimental Evaluation on Synthetic Tree

[Graphs showing comparison between MENTS and UCT with different parameters]