Maximum Entropy Monte-Carlo Planning

Chenjun Xiao[1,2], Jincheng Mei[1,2], Ruitong Huang[3], Dale Schuurmans[1,2], Martin Müller[1,2]

Entropy Regularized Value Functions

Maximum entropy policy optimization

$$\max_{\pi} \left\{ \pi \cdot \mathbf{r} + \tau \mathcal{H}(\pi) \right\}.$$

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

Note The optimal solution is given by *softmax*.

$$\mathcal{F}_\tau(\mathbf{r}) = \max_{\pi} \left\{ \pi \cdot \mathbf{r} + \tau \mathcal{H}(\pi) \right\} = \mathbf{f}_\tau(\mathbf{r}) \cdot \mathbf{r} + \tau \mathcal{H}(\mathbf{f}_\tau(\mathbf{r})).$$

where $\mathbf{f}_\tau(\mathbf{r}) = e^{(\mathbf{r} - \mathcal{F}_\tau(\mathbf{r}))/\tau}, \mathcal{F}_\tau(\mathbf{r}) = \tau \log(\mathbf{1} \cdot e^{\mathbf{r}/\tau})$.

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

Softmax policy

$$\pi^*_{\text{sft}}(a|s) = \exp\left\{ \left(Q^*_{\text{sft}}(s,a) - V^*_{\text{sft}}(s)\right)/\tau \right\}$$

Softmax value functions

$$Q^*_{\text{sft}}(s,a) = R(s,a) + \mathbb{E}_{s'|s,a}\left[V^*_{\text{sft}}(s')\right] \quad V^*_{\text{sft}}(s) = \tau \log \sum\nolimits_a \exp\left\{Q^*_{\text{sft}}(s,a)/\tau\right\}$$

Stochastic Softmax Bandit

Setup

- Sequential decision making
- K actions, (unknown) expected rewards $\mathbf{r} \in \mathbb{R}^K$
- At each round t, play A_t and receive R_t (σ^2-subgaussian)
- Goal: estimate $\mathcal{F}_\tau(\mathbf{r})$

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

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

Theorem (lower bound) For any algorithm that achieves $\mathcal{E}_t = O(\frac{1}{t})$, there exists a problem setting such that

$$\lim_{t\to\infty} t\mathcal{E}_t \geq \frac{\sigma^2}{\tau^2}\left(\mathbf{1} \cdot e^{\mathbf{r}/\tau}\right)^2.$$

Furthermore, to achieve this lower bound, there must be for any $a \in \mathcal{A}$, $\lim_{t\to\infty} N_t(a)/t = \pi^*_{\text{sft}}(a)$.

Optimal Sequential Sampling

Empirical Exponential Weight (E2W)

$$\pi_t(a) = (1-\lambda_t)\mathbf{f}_\tau(\hat{\mathbf{r}})(a) + \frac{\lambda_t}{|\mathcal{A}|}$$

where $\lambda_t = \varepsilon|\mathcal{A}|/\log(t+1)$.

Theorem E2W is asymptotically optimal, i.e.

$$\lim_{t\to\infty} t\mathcal{E}_t = \frac{\sigma^2}{\tau^2}\left(\mathbf{1} \cdot e^{\mathbf{r}/\tau}\right)^2.$$

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 called on s_T.

$$Q_{\text{sft}}(s_t, a_t) = \begin{cases} r(s_t,a_t) + R & t = T-1 \\ r(s_t,a_t) + \mathcal{F}_\tau(\mathbf{Q}_{\text{sft}}(s_{t+1})) & t < T-1 \end{cases}$$

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

$$\mathbb{P}\left\{|V_{\text{sft}}(s) - V^*_{\text{sft}}(s)| \geq \epsilon\right\} \leq \tilde{C}\exp\left(-\frac{N(s)\tau^2\epsilon^2}{C\sigma^2}\right)$$

with some constant C and $\tilde{C}$.

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

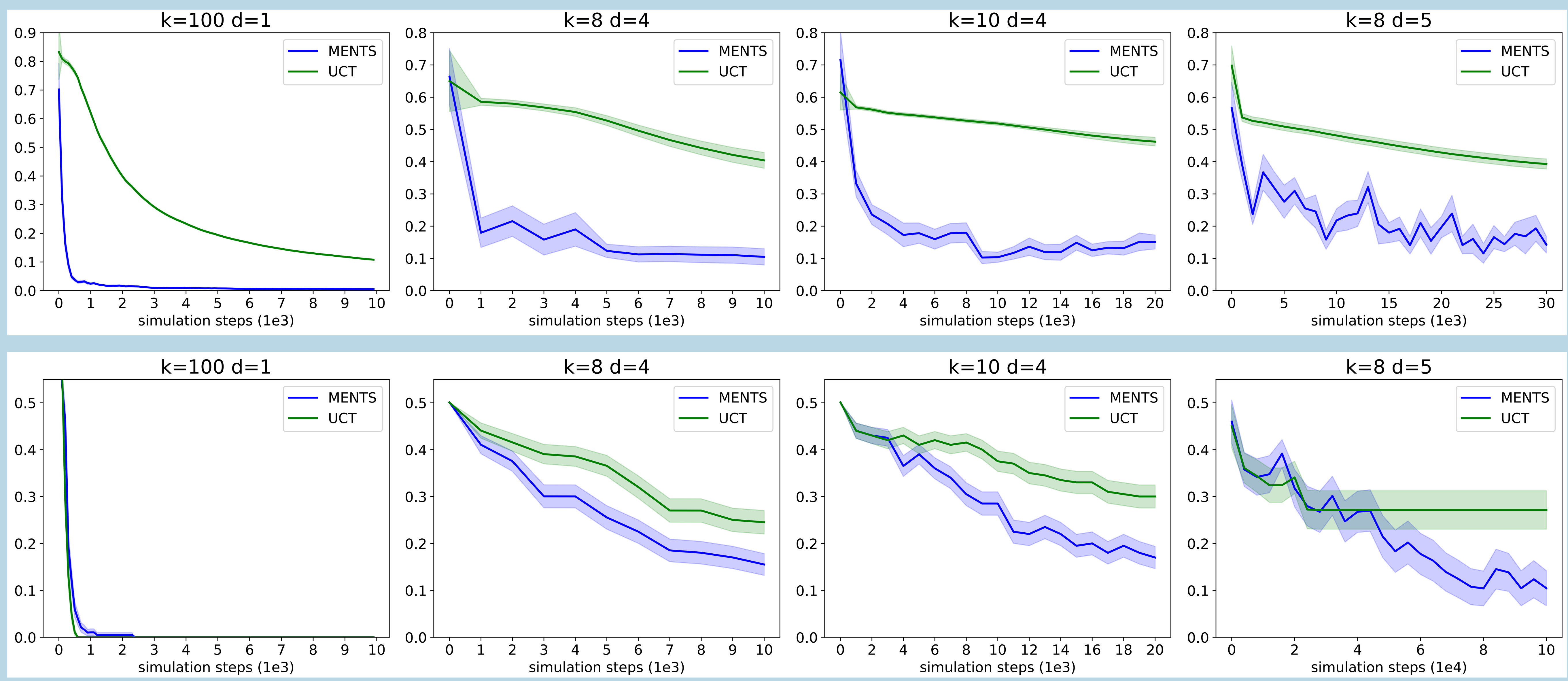
$$\mathbb{P}\left\{a_t \neq a^*\right\} \leq Ct\exp\left(-\frac{t}{(\log t)^3}\right).$$

with some constant C.

Experimental Evaluation on Synthetic Tree

value estimation

online planning



University of Alberta[1], Amii[2], BorealisAI[3]

Mailto: chenjun@ualberta.ca