Practical Open-Loop Optimistic Planning

Edouard Leurent, Odalric-Ambrym Maillard

Motivation

We consider the optimal control of an MDP $\mathcal{M} = (\mathcal{S}, A, R, T, \gamma)$ with bounded rewards $R \in [0, 1]$

↳ R and T are unknown

↳ Access to a generative model $s' \sim \mathbb{P}(s'|s, a)$ and $r \sim \mathbb{P}(r|s, a)$

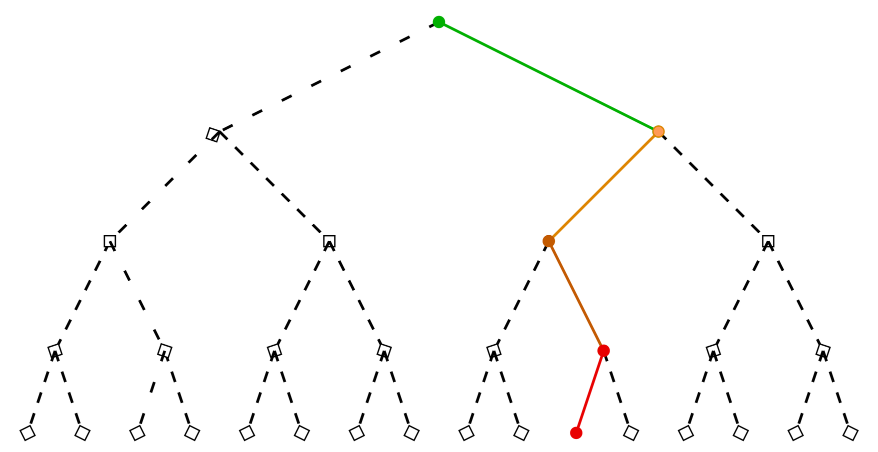
↳ Fixed-budget setting: the generative model is costly, can only be queried n times

UCT: doubly-exponential → OPD: polynomial, deterministic

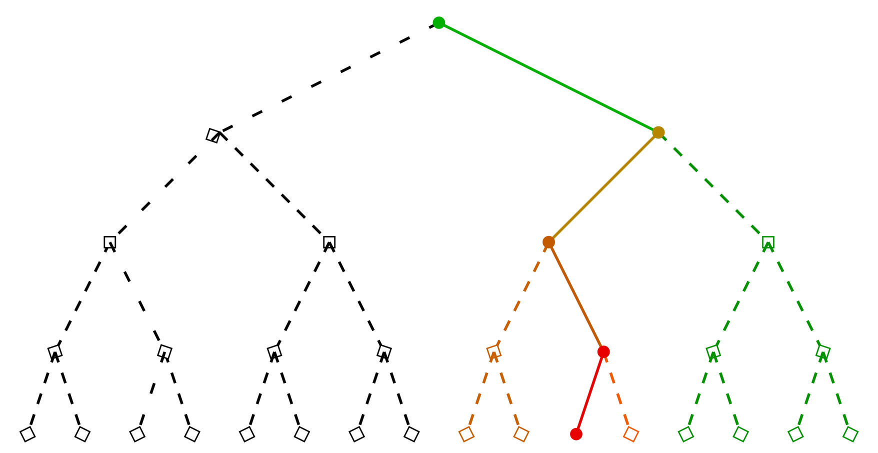
Open-Loop Optimistic Planning

OLOP algorithm introduced in [Bubeck and Munos 2010].

1. Sample M sequences of actions of fixed length L



2. Use the return structure to generalise to unseen sequences



3. Be Optimistic in the Face of Uncertainty
   ↳ in observed and future rewards

Algorithm 1: General structure for Open-Loop Optimistic Planning

1. **for** *each episode* $m = 1, \cdots, M$ **do**
2. Compute $U_a(m-1)$ from (2) for all $a \in \mathcal{T}$
3. Compute $B_a(m-1)$ from (3) for all $a \in A^L$
4. Sample a sequence with highest B-value: $a^m \in \arg\max_{a\in A^L} B_a(m-1)$.
5. **return** the most played sequence $a(n) \in \arg\max_{a\in A^L} T_a(M)$

What's wrong with OLOP?

Overly optimistic, especially in the low-budget regime.

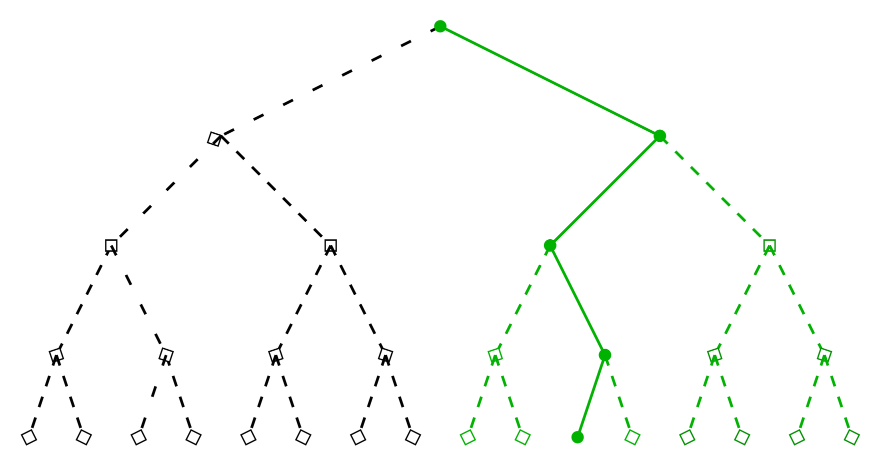
$$\underbrace{U_a^\mu(m)}_{\text{Upper bound}} \overset{\text{def}}{=} \underbrace{\hat{\mu}_a(m)}_{\text{Empirical mean}} + \underbrace{\sqrt{\frac{2\log M}{T_a(m)}}}_{\text{Confidence interval}} \quad (1)$$

$$U_a(m) \overset{\text{def}}{=} \underbrace{\sum_{t=1}^{h} \gamma^t U^\mu_{a_{1:t}}(m)}_{\text{Past rewards}} + \underbrace{\frac{\gamma^{h+1}}{1-\gamma}}_{\text{Future rewards}} \quad (2)$$

$$B_a(m) \overset{\text{def}}{=} \inf_{1\le t\le L} U_{a_{1:t}}(m) \quad (3)$$

Intuitive explanation:

- Unintended behaviour happens when $U_a^\mu(m) > 1, \forall a$.
- Then the sequence $(U_{a_{1:t}}(m))_t$ is non-decreasing
- Then $B_a(m) = U_{a_{1:1}}(m)$



OLOP behaves as uniform planning!

References

[1] Sébastien Bubeck and Rémi Munos. "Open Loop Optimistic Planning". In: *Proc. of COLT*. 2010.

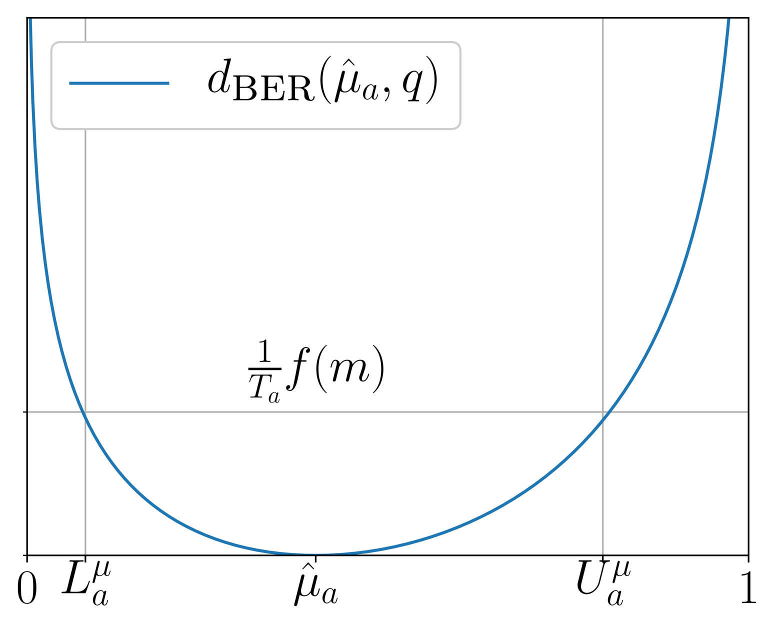
[2] Olivier Cappé et al. "Kullback-Leibler Upper Confidence Bounds for Optimal Sequential Allocation". In: *The Annals of Statistics* 41.3 (2013), pp. 1516–1541.

Kullback-Leibler OLOP

We summon the upper-confidence bound from kl-UCB [Cappé et al. 2013]:

$$U_a^\mu(m) \overset{\text{def}}{=} \max\left\{q \in I : d(\hat{\mu}_a(m), q) \le \frac{f(m)}{T_a(m)}\right\}$$

Algorithm	OLOP	KL-OLOP
Interval I	$\mathbb{R}$	[0, 1]
Divergence d	$d_{\texttt{QUAD}}$	$d_{\texttt{BER}}$
$f(m)$	$4\log M$	$2\log M + 2\log\log M$



with

$$d_{\texttt{QUAD}}(p, q) \overset{\text{def}}{=} 2(p-q)^2$$

$$d_{\texttt{BER}}(p, q) \overset{\text{def}}{=} p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$$

Conversely,

- $U_a^\mu(m) \in I = [0, 1], \forall a$.
- The sequence $(U_{a_{1:t}}(m))_t$ is non-increasing
- $B_a(m) = U_a(m)$, the *bound sharpening* step is superfluous.

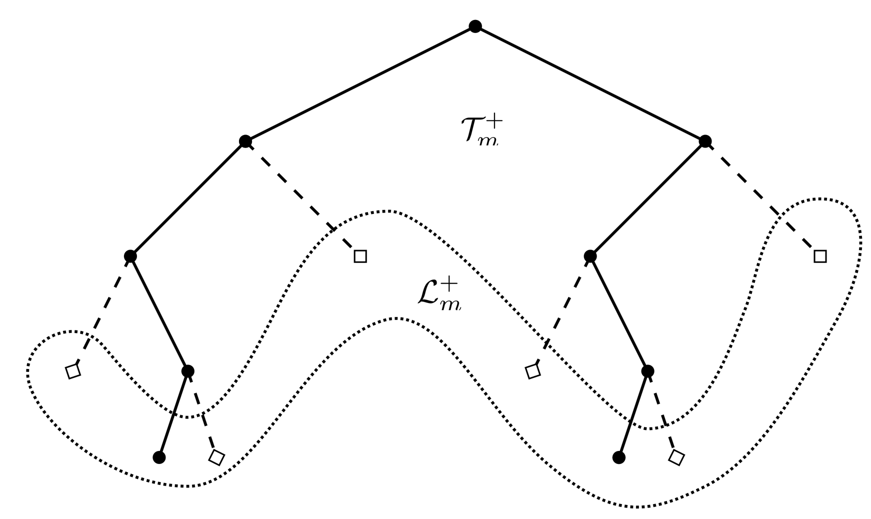
Sample complexity

Theorem 1 (Sample complexity). *KL-OLOP enjoys the same asymptotic regret bounds as OLOP. More precisely, KL-OLOP satisfies:*

$$\mathbb{E}\, r_n = \begin{cases} \tilde{0}\left(n^{-\frac{\log 1/\gamma}{\log \kappa'}}\right), & \text{if } \gamma\sqrt{\kappa'} > 1 \\ \tilde{0}\left(n^{-\frac{1}{2}}\right), & \text{if } \gamma\sqrt{\kappa'} \le 1 \end{cases}$$

Time and memory complexity

Original KL-OLOP
Compute $B_a(m-1)$ from (3) for all $a \in A^L$
Lazy KL-OLOP



Theorem 2 (Consistency). *Algorithm 2 is identical to Algorithm 1.*

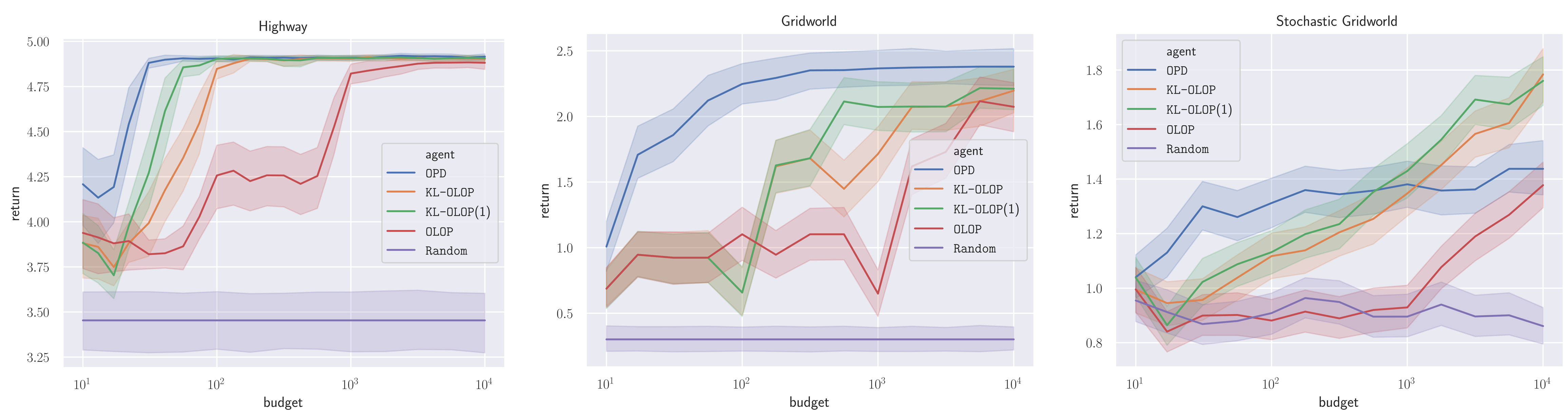
Algorithm 2: Lazy Open Loop Optimistic Planning

1. Let $\mathcal{T}_0^+ = \mathcal{L}_0^+ = \{\emptyset\}$
2. **for** *each episode* $m = 1, \cdots, M$ **do**
3. Compute $U_a(m-1)$ from (2) for all $a \in \mathcal{T}_{m-1}^+$
4. Compute $B_a(m-1)$ from (3) for all $a \in \mathcal{L}_{m-1}^+$
5. Sample a sequence with highest B-value: $a \in \arg\max_{a\in\mathcal{L}_{m-1}^+} B_a(m-1)$
6. Choose an arbitrary continuation $a^m \in aA^{L-|a|}$ // e.g. uniformly Let $\mathcal{T}_m^+ = \mathcal{T}_{m-1}^+$ and $\mathcal{L}_m^+ = \mathcal{L}_{m-1}^+$
7. **for** $t = 1, \cdots, L$ **do**
8. **if** $a_{1:t}^m \notin \mathcal{T}_m^+$ **then**
9. Add $a_{1:t-1}^m A$ to $\mathcal{T}_m^+$ and $\mathcal{L}_m^+$
10. Remove $a_{1:t-1}^m$ from $\mathcal{L}_m^+$
11. **return** the most played $a(n) \in \arg\max_{a\in\mathcal{L}_m^+} T_a(M)$

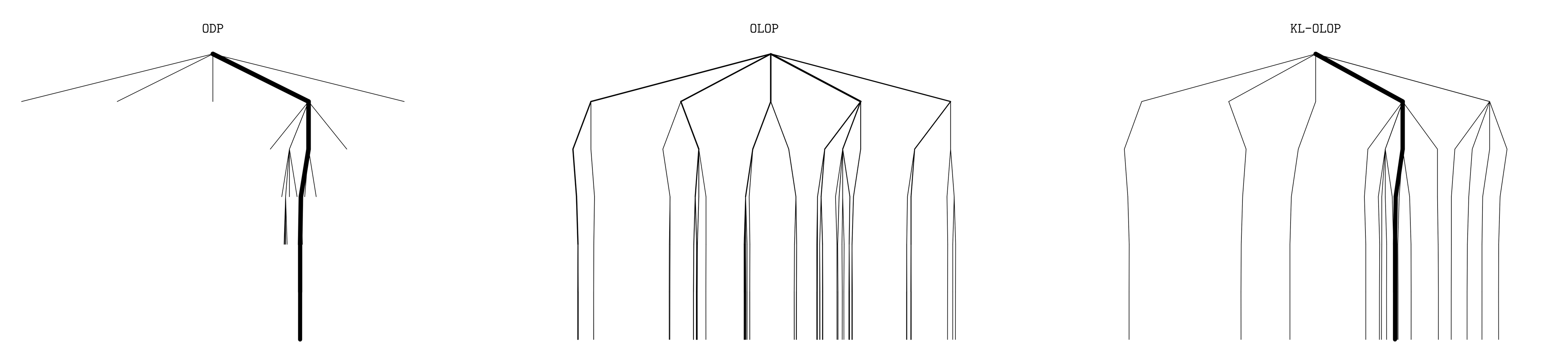
Property 1 (Time and memory complexity).

$$\frac{C(\textit{Lazy KL-OLOP})}{C(\textit{KL-OLOP})} = \frac{nA}{A^L}$$

Experiments



Average return over 100 runs — along with its 95% confidence interval — with respect to the available budget n



Expanded trees for a budget $n = 10^3$