Can we get better guarantees?

OPD: Optimistic Planning for Deterministic systems

- Introduced by [Hren and Munos 2008]
- Another optimistic algorithm
- Only for deterministic MDPs

Theorem (OPD sample complexity)

$$\mathbb{E}\, r_n = \mathcal{O}\left(n^{-\frac{\log 1/\gamma}{\log \kappa}}\right), \textit{if } \kappa > 1$$

OLOP: Open-Loop Optimistic Planning

- Introduced by [Bubeck and Munos 2010]
- Extends OPD to the stochastic setting
- Only considers open-loop policies, i.e. sequences of actions

The idea behind OLOP

A direct application of Optimism in the Face of Uncertainty

1. We want

$$\max_a V(a)$$

The idea behind OLOP

A direct application of Optimism in the Face of Uncertainty

1. We want

$$\max_a V(a)$$

2. Form upper confidence-bounds of sequence values:

$$V(a) \leq U_a \quad \text{w.h.p}$$

The idea behind OLOP

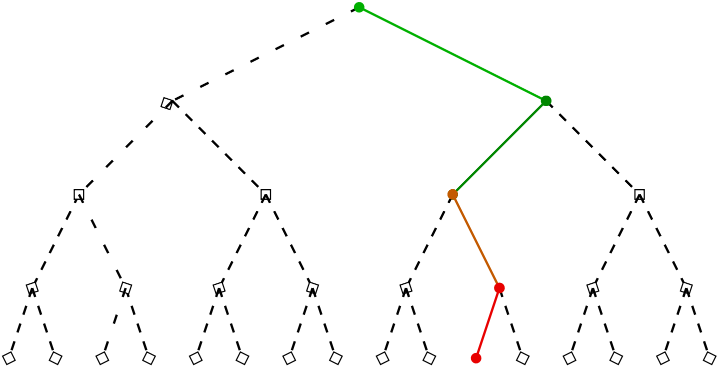
A direct application of Optimism in the Face of Uncertainty

1. We want
$$\max_a V(a)$$

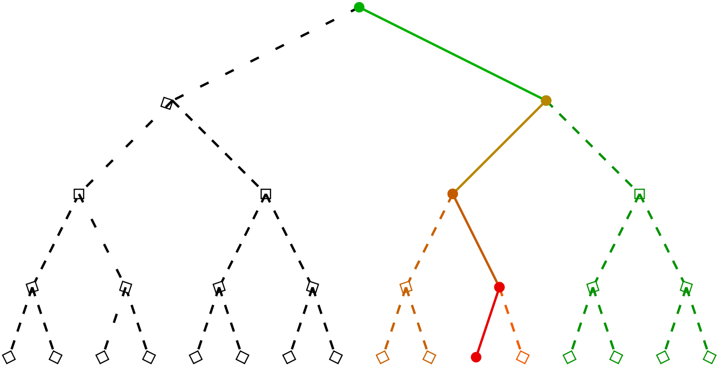
2. Form upper confidence-bounds of sequence values:
$$V(a) \leq U_a \quad \text{w.h.p}$$

3. Sample the sequence with highest UCB:
$$\arg\max_a U_a$$

The idea behind OLOP



The idea behind OLOP



Under the hood

Upper-bounding the value of sequences

$$V(a) = \overbrace{\sum_{t=1}^{h} \gamma^t \mu_{a_{1:t}}}^{\text{follow the sequence}} + \overbrace{\sum_{t \geq h+1} \gamma^t \mu_{a^*_{1:t}}}^{\text{act optimally}}$$

Under the hood

Upper-bounding the value of sequences

$$V(a) = \overbrace{\sum_{t=1}^{h} \gamma^t \underbrace{\mu_{a_{1:t}}}_{\leq U^\mu}}^{\text{follow the sequence}} + \overbrace{\sum_{t \geq h+1} \gamma^t \underbrace{\mu_{a^*_{1:t}}}_{\leq 1}}^{\text{act optimally}}$$

Under the hood

OLOP main tool: the Chernoff-Hoeffding deviation inequality

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

Under the hood

OLOP main tool: the Chernoff-Hoeffding deviation inequality

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

OPD: upper-bound all the future rewards by 1

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

Under the hood

OLOP main tool: the Chernoff-Hoeffding deviation inequality

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

OPD: upper-bound all the future rewards by 1

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

Bounds sharpening

$$B_a(m) \stackrel{\text{def}}{=} \inf_{1\leq t\leq L} U_{a_{1:t}}(m)$$

OLOP guarantees

Theorem (OLOP Sample complexity)

OLOP satisfies:

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

"Remarkably, in the case $\kappa\gamma^2 > 1$, we obtain the same rate for the simple regret as Hren and Munos (2008). Thus, in this case, we can say that planning in stochastic environments is not harder than planning in deterministic environments".