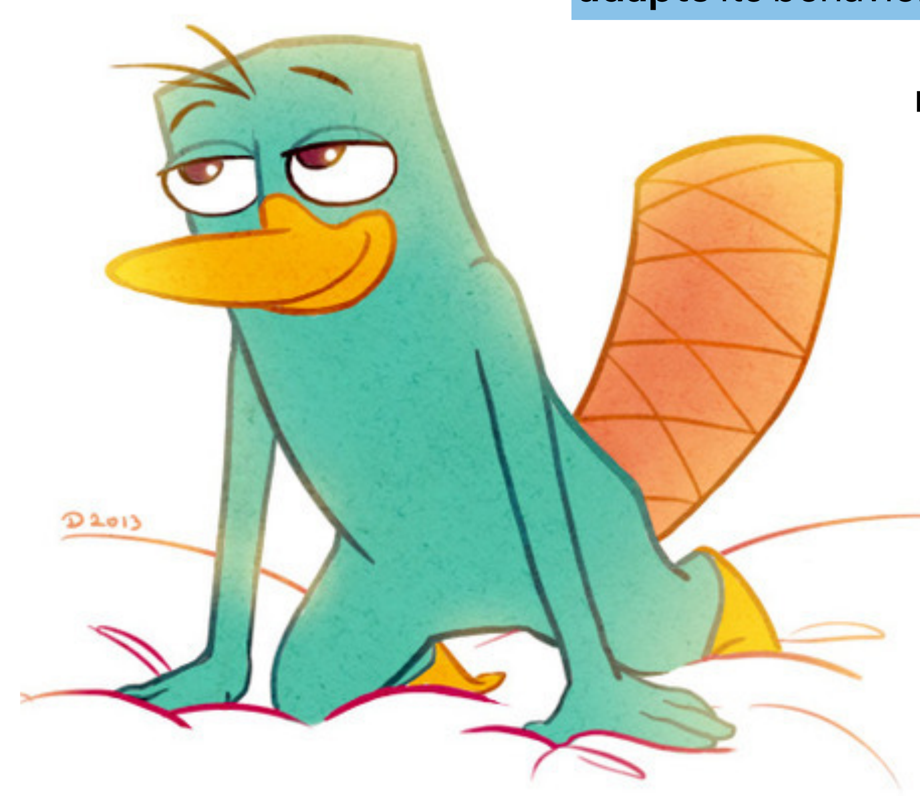
Scale-free adaptive planning for deterministic dynamics & γ-discounted rewards

Peter Bartlett, Victor Gabillon, Jennifer Healey & Michal Valko

BRAND NEW ADAPTIVE MCTS PLANNER: PlaTγPOOS

adapts its behavior to an unknown range of rewards



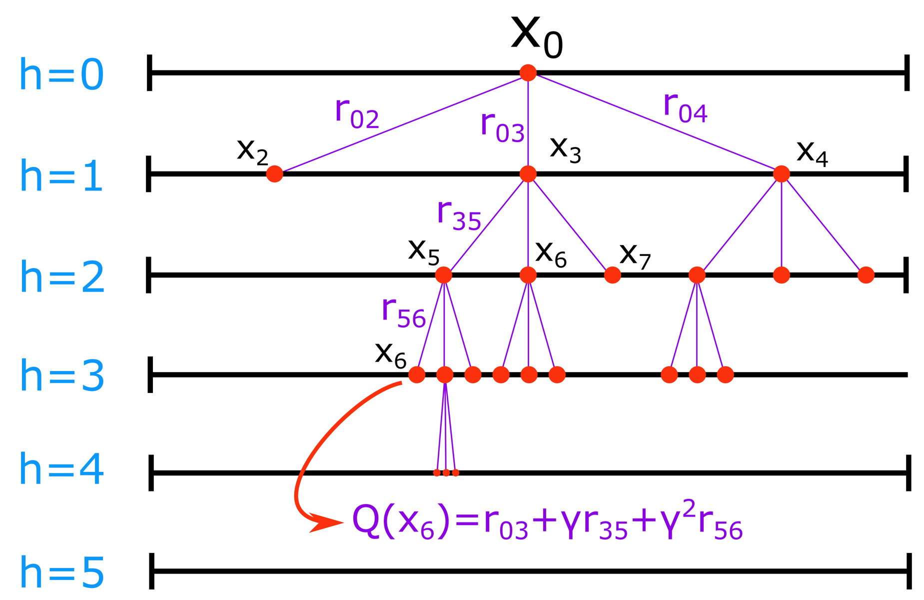
requires **no assumptions** or knowledge of noise

empirically learns much **faster** than UCB approaches

gets the **fast rate** of deterministic planning in low noise for all regimes
➡ *exponentially faster than OLOP*
not a rare case!

adapts also to the **global smoothness** ρ and beyond the base smoothness provided by **γ**

Tree search for the win!



Is this zero order optimization?

MCTS setting

MDP with **starting state** $x_0 \in X$, action space A

n **interactions:** At time t playing a_t in x_t leads to
Deterministic dynamics g: $x_{t+1} \triangleq g(x_t, a_t)$,
Reward: $r_t(x_t, a_t) + \varepsilon_t$ with ε_t being the noise

Objective: Recommend action $a(n)$ minimizing

$$r_n \triangleq \max_{a \in A} Q^\star(x,a) - Q^\star(x, a(n)) \quad \text{simple regret}$$

where $Q^\star(x,a) \triangleq r(x,a) + \sup_\pi \sum \gamma^t r(x_t, \pi(x_t))$

Assumption: $r_t \in [0, R_{\max}]$ and $|\varepsilon_t| \le b$
Approach: Explore without parameters $R_{\max}$ & b

OLOP (Bubeck and Munos, 2010)

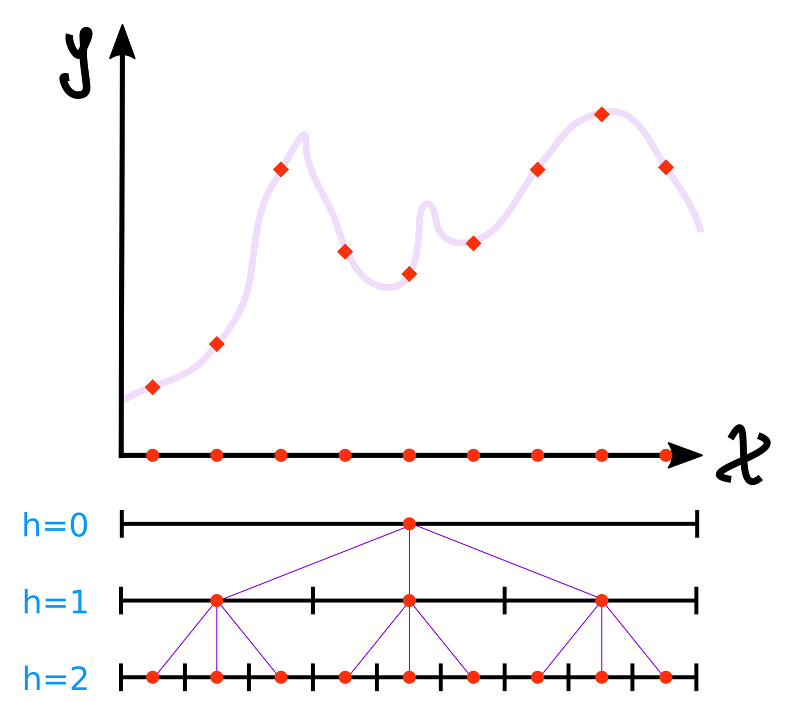
OLOP implements Optimistic Planning using Upper Confidence Bound (UCB) on the Q value of a sequence of q actions $a_1, \dots, a_q$:

$$\widehat{Q}_t(a_{1:q}) \triangleq \underbrace{\sum_{h=1}^{q}\left(\gamma^h \widehat{r}_h(t) + \frac{\gamma^h b}{\sqrt{T_{a_h}(t)}}\right)}_{\text{estimation of observed reward}} + \underbrace{\frac{R_{\max}\gamma^{q+1}}{1-\gamma}}_{\text{unseen reward}}$$

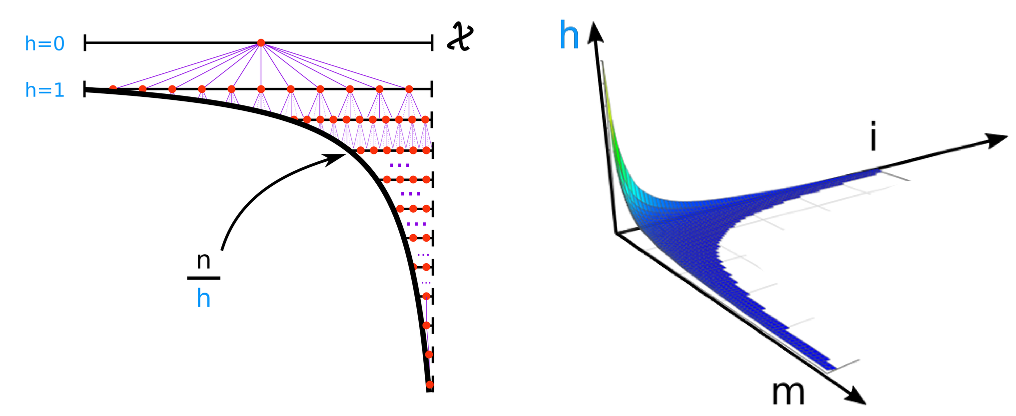
in optimization under a fixed budget n, **excellent strategies ignore** $R_{\max}$ or b

Black-box optimization

use the partitioning to explore f (uniformly)

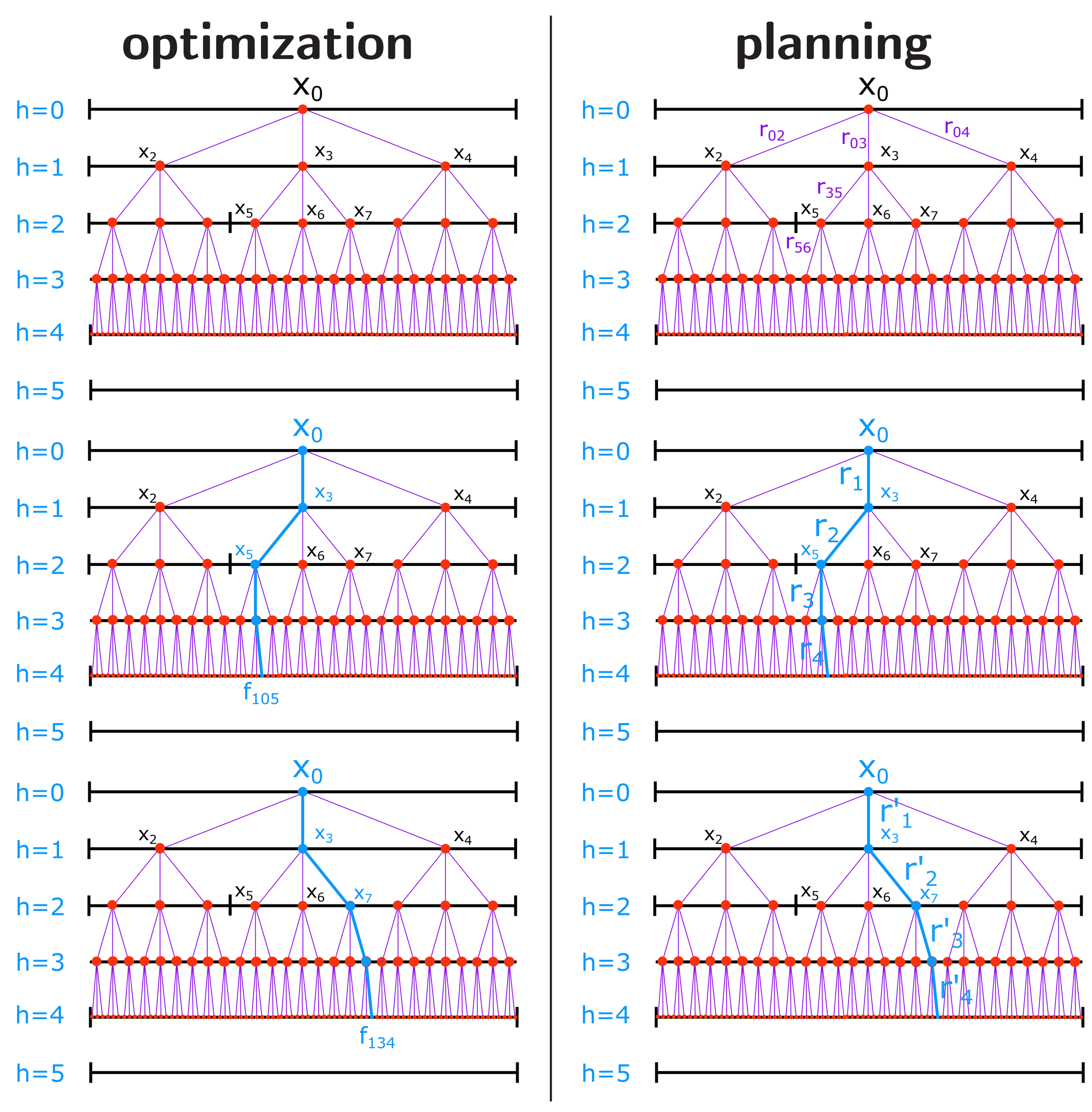


Zipf: SequOOL and StroquOOL

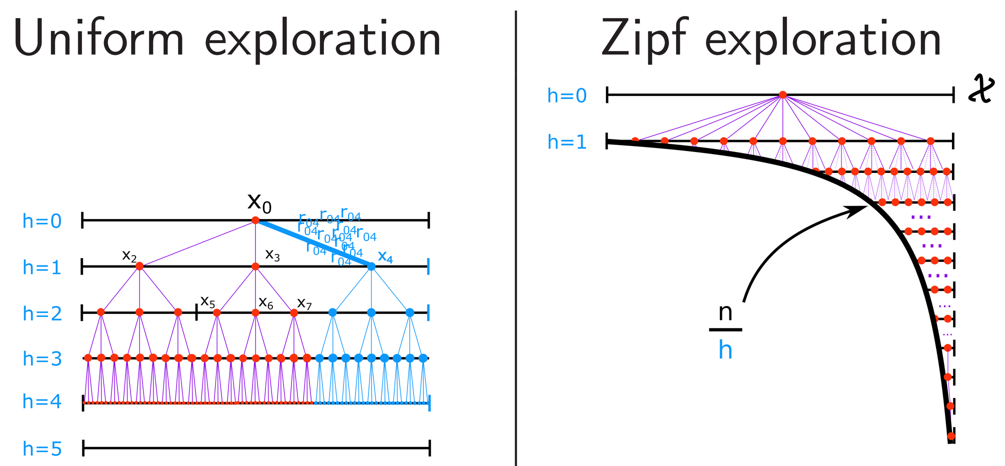


A simple parameter-free and adaptive approach to optimization under a minimal local smoothness assumption, Bartlett, Gabillon and Valko, Algorithmic Learning Theory, 2019

Optimization vs. planning



lower regret for planning! (Bubeck+Munos'10)
thanks to the **reuse of samples**



not sharing information | Sharing information

Bubeck & Munos: Only for uniform strategies ...
We figured the amount the samples needed!

Our lovely PlaTγPOOS

Input: n, A
Initialization: open the root node $\emptyset$, $h_{\max}$ times
$h_{\max} \leftarrow \left\lfloor \frac{n}{2(\log_2 n+1)^2} \right\rfloor$, $p_{\max} \leftarrow \lfloor \log_2(h_{\max}) \rfloor$

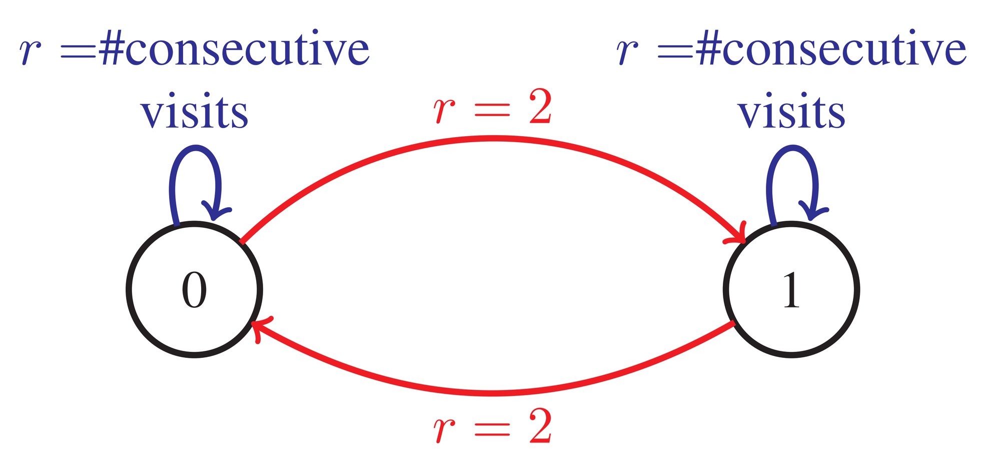
For $h = 1$ to $h_{\max}$ ◀ *exploration* ▶
For $p = \lfloor \log_2(h_{\max}/\lceil h^2\gamma^{2h}\rceil) \rfloor$ down to 0
open $\lceil h2^p\gamma^{2h}\rceil$ times the at most $\left\lfloor \frac{h_{\max}}{h\lceil h2^p\gamma^{2h}\rceil} \right\rfloor$
non-opened $a^{h,i} \in A^h$ with highest values
$\widehat{u}(a^{h,i})$ and given $T_{a^{h,i}} \ge \lceil (h-1)2^p\gamma^{2(h-1)} \rceil$

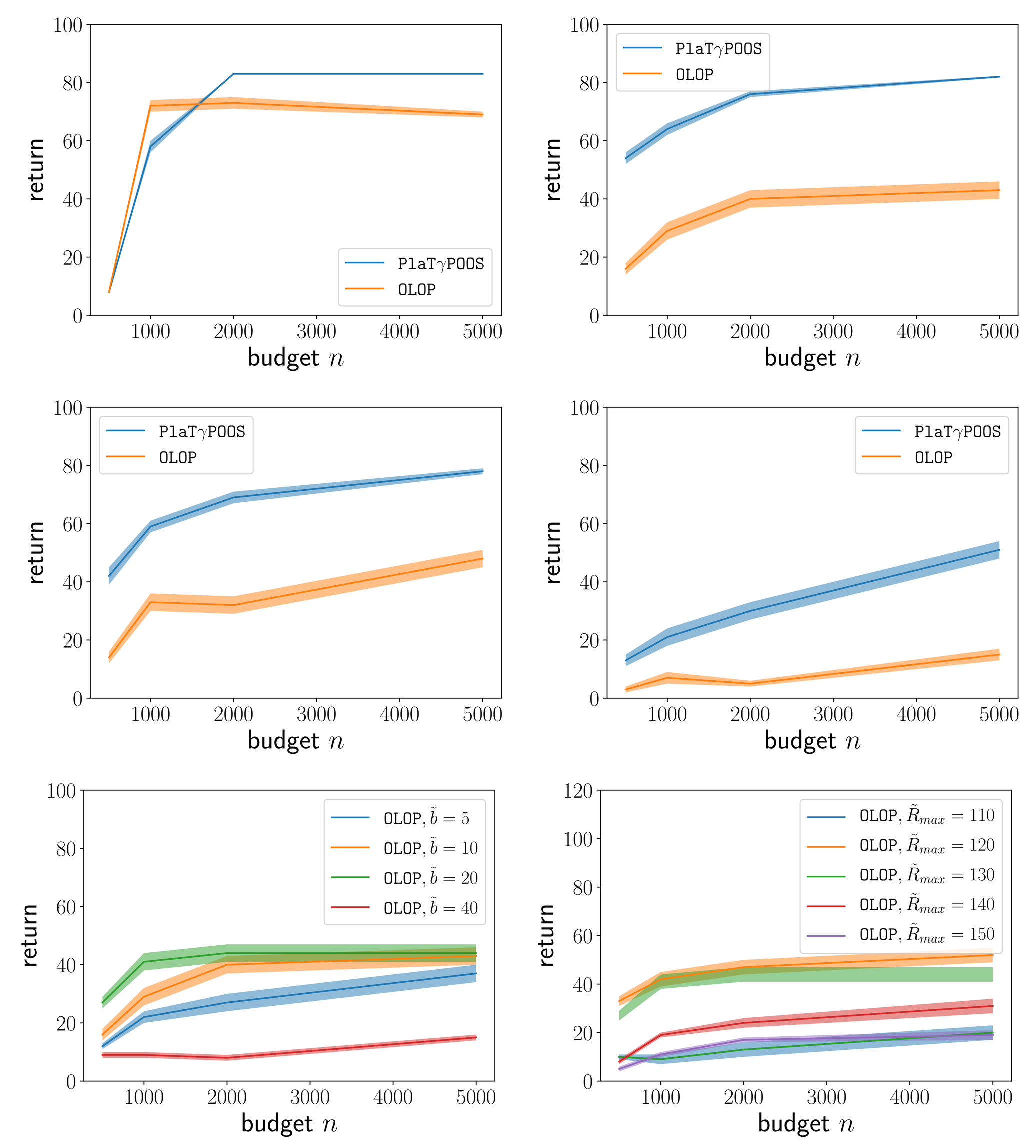
For $p \in [0 : p_{\max}]$ ◀ *cross-validation* ▶
evaluate $(t+1)\gamma^{2t}h_{\max}(1-\gamma^2)^2$ times
the actions at round t, a_t^p, of the *candidates*:
$a^p \leftarrow \underset{a \in A^\bullet : \forall t \in [2:h(a)], T_{a_{[t]}} \ge \lceil (t-1)2^p\gamma^{2(t-1)} \rceil}{\arg\max} \widehat{u}(a)$

Output $a^n \leftarrow \underset{\{a^p, p \in [0:p_{\max}]\}}{\arg\max} \widehat{u}(a^p)$

- implements **Zipf** exploration for MCTS StroquOOL
- explicitly pulls an action at depth $h+1$, γ times less than action at depth h, $(Q^\star(x,a) = r(x,a) + \sup_\pi \sum \gamma^t r(x_t, \pi(x_t))$,
- does not use UCB & no use of $R_{\max}$ and b

Numerical simulations





$b = 10$ (top center), $b = 20$, (top right), $b = 50$ (bottom left). *Bottom right:* true b is set to 10.

Empirical behavior in the figures mimics the behavior of the complexities in the table.

	$\gamma^2\kappa \le 1$		$\gamma^2\kappa \ge 1$	
	High noise (ii)	*Low noise* (ii)	*High noise* (iii)	*Low noise* (iii)
$\mathcal{E}$	$\left(\frac{n}{b^2}\right)^{-\frac{1}{2}}$	$\rho^{\sqrt{n}}$	$\left(\frac{n}{b^2}\right)^{-\frac{\log(1/\rho)}{\log(\gamma^2\kappa/\rho^2)}}$	$\left(\frac{n}{b^2}\right)^{-\frac{\log(1/\rho)}{\log(\kappa)}}$
ε	$\left(\frac{n}{b^2}\right)^{-\frac{\log(1/\rho)}{\log(\kappa)}}$	$\kappa = 1 : \rho^n$ $\kappa > 1 : \left(\frac{n}{b^2}\right)^{-\frac{\log(1/\rho)}{\log(\kappa)}}$	$\left(\frac{n}{b^2}\right)^{-\frac{\log(1/\rho)}{\log(\kappa)}}$	$\left(\frac{n}{b^2}\right)^{-\frac{\log(1/\rho)}{\log(\kappa)}}$