Monte-Carlo Graph Search: the Value of Merging Similar States

Edouard Leurent, Odalric-Ambrym Maillard

Proceedings of Machine Learning Research 129:577–602, 2020 ACML 2020

Monte-Carlo Graph Search: the Value of Merging Similar States

Edouard Leurent EDOUARD.LEURENT@INRIA.FR
Odalric-Ambrym Maillard ODALRIC.MAILLARD@INRIA.FR
Inria Lille – Nord Europe, 40 avenue Halley, 59650 Villeneuve d'Ascq, France

Editors: Sinno Jialin Pan and Masashi Sugiyama

Abstract

We consider the problem of planning in a Markov Decision Process (MDP) with a generative model and limited computational budget. Despite the underlying MDP transitions having a graph structure, the popular Monte-Carlo Tree Search algorithms such as `UCT` rely on a tree structure to represent their value estimates. That is, they do not identify together two similar states reached via different trajectories and represented in separate branches of the tree. In this work, we propose a *graph-based* planning algorithm, which takes into account this state similarity. In our analysis, we provide a regret bound that depends on a novel problem-dependent measure of difficulty, which improves on the original tree-based bound in MDPs where the trajectories overlap, and recovers it otherwise. Then, we show that this methodology can be adapted to existing planning algorithms that deal with stochastic systems. Finally, numerical simulations illustrate the benefits of our approach.
Keywords: Online Planning, Tree-search, Reinforcement Learning.

1. Introduction

Monte Carlo tree search (MCTS) algorithms (Coulom, 2006) were a breakthrough for online decision-making in Markov decision processes (MDPs), that lead to key successes in the domain, including Computer Go (Silver et al., 2018). They enjoy two main benefits: first, they do not require the knowledge of the MDP parameters contrary to *e.g.* Dynamic Programming algorithms, but only the access to a *generative model* that allows to sample trajectories from the current state. Second, the theoretical performance bounds of MCTS algorithms are typically independent of the size of the state space. Instead, they depend on the maximum depth at which an optimal node in the search tree can be reached within the allowed *budget* of trajectory samples. This translates as an *effective branching factor* in the bounds, related to the notion of near-optimality dimension in multi-armed bandits.

Algorithms for planning with a generative model date back at least to the seminal work of Kearns et al. (2002) who proposed the `Sparse Sampling` algorithm using a tree structure to represent the value estimate and uniform sampling of trajectories. This strategy was further analysed more recently in (Feldman and Domshlak, 2014), where the `BRUE` algorithm provides an enhanced value estimation. Another family of algorithms rely on the principle of *Optimism in the Face of Uncertainty* (surveyed by Munos, 2014), inspired from the Multi-Armed Bandit problem. This principle was first used in the context of planning in the `CrazyStone` software (Coulom, 2006) for computer Go. It was later formalized with

the UCT algorithm (Kocsis and Szepesvári, 2006), but was shown by Coquelin and Munos (2007) to have a doubly-exponential complexity in the worst case. The Optimistic Planning for Deterministic Systems (OPD) algorithm introduced by Hren and Munos (2008) was the first to provide a polynomial regret bound, but was limited to systems with deterministic rewards and dynamics. It was then extended to the case of stochastic rewards (Bubeck and Munos, 2010; Leurent and Maillard, 2019) with deterministic transitions. Known stochastic transitions were handled by Busoniu and Munos (2012). For MDPs with stochastic and unknown transitions, polynomial sample complexities have been obtained for StOP (Szorenyi et al., 2014), TrailBlazer (Grill et al., 2016) and SmoothCruiser (Grill et al., 2019), but despite their theoretical merits these algorithms are intractable in practice: StOP requires the expensive storage of policies, while TrailBlazer and SmoothCruiser only terminate after a prohibitive amount of samples, even for very small MDPs. (Huang et al., 2017; Kaufmann and Koolen, 2017) proposed two algorithms for planning in a maxmin game with stochastic rewards in the leaves of a known game tree. The latter was recently extended by Jonsson et al. (2020) with the MDP-GapE algorithm for planning in unknown stochastic MDPs, which enjoys a gap-dependent sample complexity.

Despite its simplicity, the use of a tree structure comes with a limitation: MCTS algorithms *do not merge information across states*. That is, if a state s can be reached via two trajectories, it will be represented twice in the look-ahead tree. For instance, in Figure 1 (left), two paths lead to the same state represented in orange. MCTS algorithms do not merge the information of the two trajectories to update a shared estimate of the state value.

Related work The idea of merging information between branches of a search tree appears in (Silver et al., 2018), where the state values are approximated with a shared Neural Network. However, this network is merely updated between two planning instances and not during the planning procedure itself. Another work of interest is that of Hostetler et al. (2014), who propose to partition the state space S into a smaller set $\mathcal{X}$ of equivalence classes. By aggregating similar states within a class, they reduce the branching factor of the search tree from $|S||A|$ to $|\mathcal{X}||A|$, which substantially improves sample complexity as they illustrate empirically. However, this procedure requires providing a relevant state partition, only aggregates trajectories that traverse the same sequence of classes (*i.e.* local deformations), and comes with a (bounded) loss of optimality. The closest work to ours is that of Ballesteros et al. (2013), in the context of partially observable MDPs, who identify similar belief states and plan with a graph structure. This work focuses on empirically comparing various similarity measures on robotic tasks and does not provide any theoretical analysis of the effect of aggregation. This is precisely our goal and contribution here.

Contributions We introduce a planning algorithm named GBOP-D, a graph-based version of the tree-based OPD algorithm for deterministic systems. We analyse the benefits of this graph-based formulation in Section 4, and provide in Theorem 16 a regret guarantee. The corresponding regret bound features a novel problem-dependent difficulty measure that we introduce to capture the benefit of using a graph structure. We show that this measure can only improve over the performance of OPD, and provide an example where it does. We discuss in Section 5 an extension of our method to stochastic MDPs, called GBOP. Finally, Section 6 illustrates the benefits of GBOP in two numerical simulations.

2. Background

In a *Markov Decision Process* (MDP), an agent observes its current state s from a state space S and picks an action a from an action space A of size K, before transitioning to a next state s' drawn from a transition distribution $P(s' \mid s, a)$ and receiving a bounded reward $r \in [0, 1]$ drawn from a reward distribution $P(r \mid s, a)$. The goal of the agent is to maximise in expectation its cumulative discounted rewards $\sum_{t=0}^{\infty} \gamma^t r_t$, where $\gamma \in (0, 1)$ is a discount factor. This amounts to choosing at each step the action that maximises the state-action value function $Q(s, a) \overset{def}{=} \max_\pi \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty} \gamma^t r_t \;\middle|\; s_0 = s, a_0 = a\right]$ where π is a policy mapping states to actions and $\tau = (s_0, a_0, \dots)$ is a trajectory generated by following π. In *Monte-Carlo planning*, the underlying MDP is *unknown* to the agent which only has access to a *generative model* that provides samples s' of $P(s' \mid s, a)$ when queried. Then, under computational constraints known as *fixed-budget*, the agent is only allowed a limited budget n of queries to the generative model before recommending a good action a_n to take next. The quality of that action is assessed in terms of the *simple regret*

$$r_n \overset{def}{=} V(s) - Q(s, a_n), \text{ where } V(s) \overset{def}{=} \max_a Q(s, a). \tag{1}$$

3. Graph-Based Planning for Deterministic Systems

In this section, we introduce a simple yet highly effective variant of tree-based planning algorithms. We first consider the simple setting of MDPs with deterministic dynamics and rewards, and will denote $r(s, a)$ the unique reward r sampled from $P(r|s, a)$ and $P(s, a)$ the unique next state s' sampled from $P(s'|s, a)$. We start by giving some background on the interplay of data structures and optimistic planning algorithms.

3.1. Data structures

In this work, we compare two data structures for planning in an MDP: tree and (directed) graph, represented in Figure 1. In order to distinguish them, we referring to trees with Roman symbols, e.g. T, U, L, B; and to graphs with calligraphic symbols, e.g. $\mathcal{G}, \mathcal{U}, \mathcal{L}, \mathcal{B}$. In both structures, we say that a node is *internal* if it has outgoing edges, and *external* else.

In a tree, a node of depth h represents a sequence of actions $a \in A^h$. The *root* of the tree corresponds to the empty action sequence, and hence to the initial state $s_0 \in S$. At iteration n, we denote the current tree as T_n. Borrowing notations from topology, we denote its set of internal nodes as $\mathring{T}_n$ and its set of external nodes (the leaves) as ∂T_n. Note that since the MDP is deterministic, a sequence of action a is associated with its final state denoted $s(a)$, but this association is not one-to-one: several sequences of action can lead to the same state, which will be represented several times in the tree.

In a graph, the nodes represent states $s \in S$, and the edges represent transitions between states. The *source* of the graph corresponds to the initial state s_0. At iteration n, we denote the current graph as $\mathcal{G}_n$, its set of internal nodes as $\mathring{\mathcal{G}}_n$ and its set of *sinks* as $\partial\mathcal{G}_n$.

Both structures are built iteratively from a single starting node, by selecting an external node (leaf or sink) to expand. The *expansion* of a node a or s refers to calling the generative model to sample the reward r and next state s' for each action $a \in A$, and adding child nodes to the data structure. In a tree, the expansion of a node $a \in A^h$ always lead to the

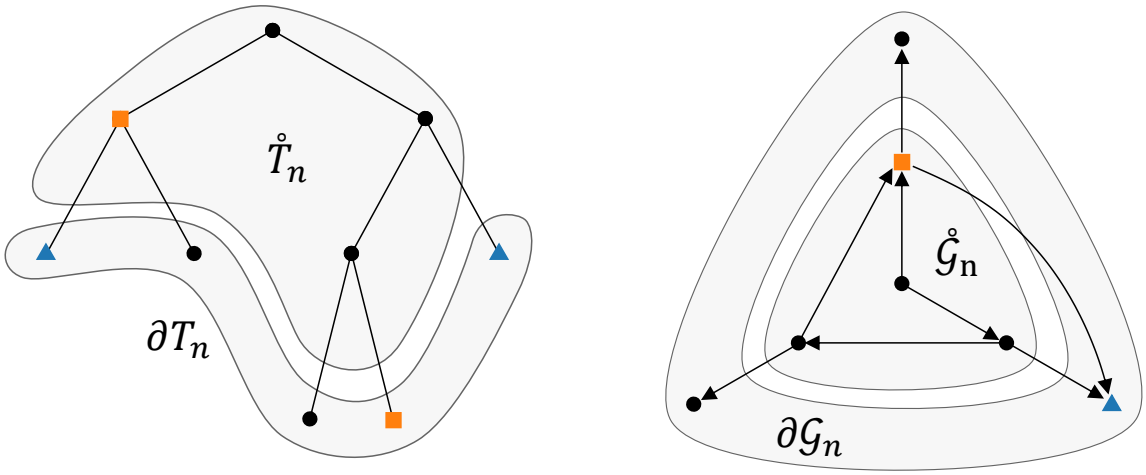


Figure 1: Illustration of the tree T_n (left) and the graph $\mathcal{G}_n$ (right) built from the same observed transitions. The root of the tree corresponds to the central graph node. In the tree, two nodes with the same colour and shape (not black round) lead to the same state.

creation of new leaves that represent the suffix sequence of action $ab \in A^{h+1}$, $b \in A$. The maximum depth of an expanded node in T_n is denoted d_n. In contrast, in a graph the next state s' reached from s, a might already be present in $\mathcal{G}_n$, in which case we add the edge between s and s' without creating a new node. These data structures can be used to store information about the MDP, such as the transitions and rewards $r(s, a)$, or other informations useful for planning.

3.2. Optimistic planning

A planning algorithm is typically composed of two main rules: (i) A *sampling rule*, that selects promising transitions to simulate at each iteration n; (ii) A *recommendation rule*, that recommends a good first action a_n to take (in s_0). These rules can be chosen with the goal of minimising the simple regret r_n. A popular approach is to follow the principle of *Optimism in the Face of Uncertainty* (OFU) (see Munos, 2014), which consists in exploring the option that maximises an upper-bound of the true objective. In the context of planning, it has been applied by forming bounds on the value function V.

Definition 1 (Value bounds) ***On trees.*** *We denote by $L : T_n \to \mathbb{R}$ and $U : T_n \to \mathbb{R}$ a lower-bound and upper-bound for the state value V defined on the tree T_n, such that*

$$\forall a \in T_n, \qquad L(a) \leq V(s(a)) \leq U(a).$$

In the sequel, we abuse notations and denote this inequality as $L \leq V \leq U$.

On graphs. *Likewise, we denote by $\mathcal{L} : \mathcal{G}_n \to \mathbb{R}$ and $\mathcal{U} : \mathcal{G}_n \to \mathbb{R}$ a lower-bound and upper-bound for the state value V defined on the graph $\mathcal{G}_n$, such that*

$$\forall s \in \mathcal{G}_n, \qquad \mathcal{L}(s) \leq V(s) \leq \mathcal{U}(s).$$

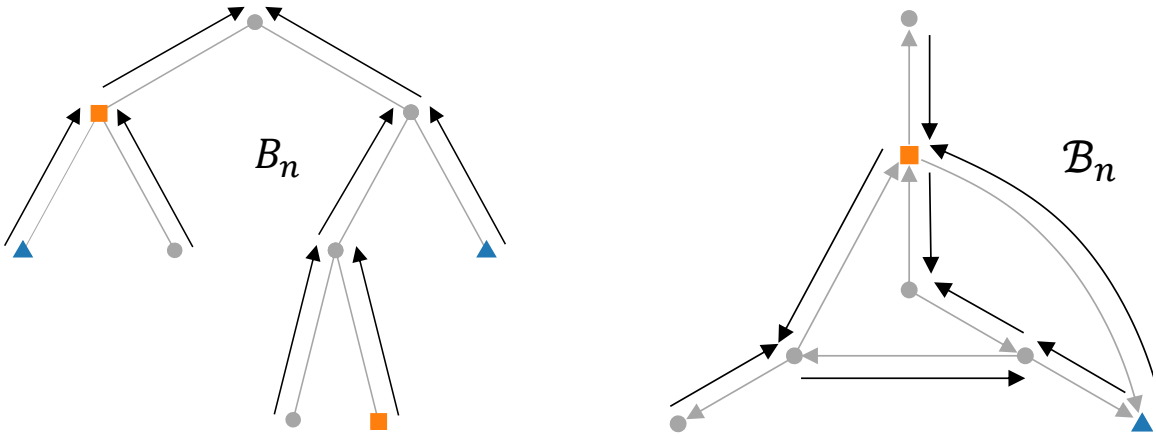


Figure 2: Black arrows depict how the Bellman backup operators B_n (left) and $\mathcal{B}_n$ (right) propagate value estimates from successor nodes to their parents. Information travels freely in a graph, but only upwards in a tree.

Following the OFU principle, at iteration n we must leverage available information to design an upper-bound U_n (or $\mathcal{U}_n$) on V as tight as possible. Then, in order to select a promising external node to expand, the sampling rule starts from the root (or source) and follows the optimistic strategy of always selecting the action which maximises U_n (or $\mathcal{U}_n$), until reaching an optimistic leaf (or sink) to expand. This strategy was used with great success in (e.g. Kocsis and Szepesvári, 2006; Hren and Munos, 2008; Bubeck and Munos, 2010; Busoniu and Munos, 2012).

For instance, since we assume that the rewards are bounded in $[0, 1]$, trivial bounds on $V(s)$ are $0 \leq V(s) \leq V_{\max} \overset{def}{=} \sum_t \gamma^t 1 = 1/(1-\gamma)$. But these trivial bounds are the same for every node, which makes them non-informative, and do not make use of the observed information. However, they can be used as a valid starting point. Each observed transition can then be used to tightened these bounds, by resorting to the Bellman optimality operator.

Definition 2 (Bellman optimality operator) ***On trees.*** *We define the Bellman optimality operator B_n on the tree T_n as:*

$$B_n(f)(a) \overset{def}{=} \begin{cases} \max_{b \in A} r(s(a), b) + \gamma f(ab) & \text{if } a \in \mathring{T}_n; \\ f(a) & \text{if } a \in \partial T_n; \end{cases} \tag{2}$$

where $f : T_n \to \mathbb{R}$ is a real-valued function of tree nodes, such as the value bounds L, U.

On graphs. *Likewise, we define the Bellman optimality operator $\mathcal{B}_n$ on the graph $\mathcal{G}_n$:*

$$\mathcal{B}_n(f)(s) \overset{def}{=} \begin{cases} \max_{b \in A} r(s, b) + \gamma f(P(s, b)) & \text{if } s \in \mathring{\mathcal{G}}_n; \\ f(s) & \text{if } s \in \partial \mathcal{G}_n. \end{cases} \tag{3}$$

The updates with both Bellman operators are depicted in Figure 2.

Hren and Munos (2008) used this Bellman operator B_n in their `OPD` algorithm to define a pair of bounds (L_n, U_n) at each iteration n. They use trivial bounds at the leaves, and

backup these estimates up to the root by iteratively applying B_n. We can show that, under a *monotonicity* condition (satisfied by the trivial bounds 0 and V_{max}), applying B_n can only tighten a bound and converges in a finite time.

Definition 3 (Monotonicity) *A pair of bounds (L, U) or $(\mathcal{L},\mathcal{U})$ is* monotonic *if they are respectively non-decreasing and non-increasing along transitions:*

$$\begin{aligned} \forall a \in T_n, \quad & L(a) \leq B_n(L)(a), && U(a) \geq B_n(U)(a) \\ \forall (s) \in \mathring{\mathcal{G}}_n, \quad & \mathcal{L}(s) \leq \mathcal{B}_n(\mathcal{L})(s), && \mathcal{U}(s) \geq \mathcal{B}_n(\mathcal{U})(s) \end{aligned}$$

Lemma 4 (Properties of B_n) *(i) B_n preserves monotonicity and tightens monotonic bounds:*

$$\text{if } L \leq V \leq U, \text{ then } L \leq B_n(L) \leq V \leq B_n(U) \leq U;$$

(ii) The sequence $B_n^k = \underbrace{B_n \circ \cdots \circ B_n}_{k \text{ times}}$ converges in a finite time $k = d_n$, where d_n is the depth of T_n.

This enables Hren and Munos (2008) to define[1] non-trivial valid bounds on V:

$$L_n \overset{def}{=} B_n^{d_n}(0), \qquad U_n \overset{def}{=} B_n^{d_n}(V_{\max}), \tag{4}$$

where 0 is the null function. The corresponding `OPD` algorithm is described in Algorithm 1.

Algorithm 1: The *Optimistic Planning of Deterministic Systems* (`OPD`) algorithm from (Hren and Munos, 2008).

for *each iteration n* **do**
  Compute the bounds $L_n = B_n^{d_n}(0)$ and $U_n = B_n^{d_n}(V_{\max})$.
  $b_n \leftarrow \emptyset$
  while *the node $b_n \in \mathring{T}_n$ is internal* **do**
    $b_n \leftarrow \arg\max_{a' \in b_n A} r(a') + \gamma U_n(a')$ ▷ `Optimistic sampling rule`
  for *action $a \in A$* **do** ▷ `Node expansion`
    Simulate $r \leftarrow r(s(b_n), a)$ and $s' \leftarrow P(s(b_n), a)$.
    Add a new leaf $b_n a$ to T_{n+1}, with associated reward r.
return $\arg\max_{a \in A} r(s,a) + \gamma L_n(a)$. ▷ `Conservative recommendation rule`

Likewise, we show that the graph version $\mathcal{B}_n$ verifies similar properties.

Lemma 5 (Properties of $\mathcal{B}_n$) *(i) $\mathcal{B}_n$ preserves monotonicity and tightens monotonic bounds:*

$$\text{if } \mathcal{L} \leq V \leq \mathcal{U}, \text{ then } \mathcal{L} \leq \mathcal{B}_n(\mathcal{L}) \leq V \leq \mathcal{B}_n(\mathcal{U}) \leq \mathcal{U};$$

(ii) $\mathcal{B}$ is a γ-contraction, and we denote $\mathcal{B}_n^\infty \overset{def}{=} \lim_{k\to\infty} \mathcal{B}_n^k$.

This motivates us to propose Algorithm 2, following the approach of Algorithm 1 adapted to a graph structure.

1. We use an iteration of operators while a recursive definition was used originally.

Algorithm 2: Our proposed *Graph-Based Optimistic Planning for Deterministic systems* (GBOP-D) algorithm.

for *each iteration* n **do**

1 Compute the bounds $\mathcal{L}_n = \mathcal{B}_n^\infty(0)$ and $\mathcal{U}_n = \mathcal{B}_n^\infty(V_{\max})$.

$s_n \leftarrow s_0$

while *the node* $s_n \in \mathring{\mathcal{G}}_n$ *is internal* **do**

2 $s_n \leftarrow \arg\max_{s'} r(s_n, a) + \gamma \mathcal{U}_n(s')$ ▷ `Optimistic sampling rule`

for *action* $a \in A$ **do** ▷ `Node expansion`

Simulate $r \leftarrow r(s_n, a)$ and $s' \leftarrow P(s_n, a)$.

Get or create the node s' in $\mathcal{G}_{n+1}$, and add the transition $(s_n, a) \to s', r$.

return $\arg\max_{a \in A} r(s, a) + \gamma \mathcal{L}_n(s(a))$. ▷ `Conservative recommendation rule`

Remark 6 (Termination and complexity) *One key difficulty both in the design and analysis of the algorithm is the correct handling of loops in the graph. Indeed, there are two procedures in GBOP-D that may not terminate in finite time whenever $\mathcal{G}_n$ contains a loop: the computation of $\mathcal{B}_n^\infty$ (line 1) and the sampling rule loop (line 2). We handle these steps carefully in the Supplementary Material, where we discuss an approximate implementation in which these two procedures are stopped whenever they reach a desired accuracy ε, along with an analysis of the corresponding time complexity and impact on the performance.*

Though both algorithms share a similar design, we claim that using graphs provides substantial theoretical and practical performance improvements, and back up this statement in Sections 4 and 6.

4. Analysis

Comparing OPD and GBOP-D directly is difficult since they do not involve the same structure, which implies implicit differences in their behaviours. Studying them under a common framework makes these differences explicit. To leverage the analysis of OPD by Hren and Munos (2008), we will frame GBOP-D as a tree-based planning algorithm: the graph operator $\mathcal{B}$ will be represented as tree backup B applied on an *unrolled* tree $T(\mathcal{G}_n)$, defined below.

4.1. Background on the sample complexity of OPD

First, we recall the analysis of Hren and Munos (2008) and introduce some notations.

Definition 7 (Sequence values) *The value of a finite* ***sequence*** *of actions $a \in A^h$ is:*

$$V(a) = R(s_0, a) + \gamma^h V(s(a)),$$

where $R(s, a) = \sum_{t=0}^{h-1} \gamma^t r_t$ is the return obtained by executing the sequence of actions a starting from the state s. We also denote $V^\star = \max_{a \in A} V(a)$ the value of the best action.

This enables to define a measure of the difficulty of a planning problem.

Definition 8 (Difficulty measure) *We define the near-optimal branching factor* $\boldsymbol{\kappa}$ *of an MDP as*

$$\boldsymbol{\kappa} = \limsup_{h\to\infty} |\boldsymbol{T}_h^\infty|^{1/h} \in [1, K] \tag{5}$$

where $T_h^\infty = \left\{ a \in A^h : V^\star - V(a) \leq \frac{\gamma^h}{1-\gamma} \right\}$ *is the set of near-optimal nodes at depth* h.

This problem-dependent measure $\boldsymbol{\kappa}$ is the branching factor of the subtree $T^\infty = \bigcup_h T_h^\infty$ of near-optimal nodes that can be sampled by OPD, and acts as an effective branching factor as opposed to the true branching factor K. When $\boldsymbol{\kappa}$ is small, fewer nodes must be explored at a given depth allowing the algorithm to plan deeper for a given budget n. Thus, it directly impacts the simple regret that can be achieved by OPD when run on a given MDP.

Theorem 9 (Regret bound of Hren and Munos 2008) *The Algorithm 1 enjoys the following regret bound:*

$$r_n = \widetilde{\mathcal{O}}\left(n^{-\log\frac{1}{\gamma}/\log\boldsymbol{\kappa}}\right),$$

where $f_n = \widetilde{\mathcal{O}}(n^{-\alpha})$ *means that for any* $\alpha' < \alpha$, $f_n = \mathcal{O}(n^{-\alpha'})$, *for all* $\alpha \in \mathbb{R}_+ \cup \{+\infty\}$.

The near-optimal branching factor $\boldsymbol{\kappa}$ is related (Bubeck and Munos, 2010) to the near-optimality dimension studied in the online optimisation literature (see e.g. Bubeck et al., 2009; Munos, 2011). It is typically small in problems where there is one single optimal trajectory, of which any deviation can be quickly dismissed as suboptimal. Conversely, κ is large when many sub-optimal trajectories cannot be distinguished easily based on their values, which requires the exploration of a large part of the tree T of branching factor K.

4.2. Motivation for an improved regret bound

We start by reformulating the sampling rule used for the OPD algorithm. To that end, notice that when some bounds (L, U) on the state values $V(s(a))$ are available, they also induce bounds $(\overline{L}, \overline{U})$ on values $V(a)$ of sequences of actions a of length h defined as:

$$\underbrace{R(s_0,a) + \gamma^h L(a)}_{\overline{L}(a)} \leq V(a) \leq \underbrace{R(s_0,a) + \gamma^h U(a)}_{\overline{U}(a)}.$$

One can easily see that, since the (L_n, U_n) used in the optimistic sampling rule described in Algorithm 1 are invariant by B_n by definition, this rule can be equivalently expressed as:

$$b_n \in \arg\max_{a\in\partial T_n} \overline{U}_n(a). \tag{6}$$

Likewise, the conservative recommendation rule returns the first action of:

$$a_n \in \arg\max_{a\in\partial T_n} \overline{L}_n(a) \tag{7}$$

As shown in Figure 2, in a tree the Bellman operator B_n only propagates the information upward, and the leaves cannot be updated. Thus, $U_n = B_n^{d_n}(V_{\max})$ and $V_{\max}$ coincide on

∂T_n which means that the sampling rule of OPD can be summarized as using (6) with the trivial upper-bound $U_n = V_{\max}$. Likewise, the recommendation rule simply uses (7) with the trivial lower-bound $L_n = 0$. Thus, OPD amounts to simply using the trivial bound $(0, V_{\max})$ on leaf nodes, and does not make use of all the available information in T_n to improve these bounds.

Let us now assume for the moment that we have access to tighter bounds (L, U) provided by an oracle:

$$0 \leq L \leq V \leq U \leq V_{\max}.$$

Definition 10 (A finer difficulty measure) *We define the near-optimal branching factor* according to the bounds (L, U) *as*

$$\boldsymbol{\kappa(L,U) \overset{def}{=} \limsup_{h\to\infty} \left|T_h^\infty(L,U)\right|^{1/h} \in [1,K],} \tag{8}$$

where $T_h^\infty(L,U) = \left\{a \in A^h : V^\star - V(a) \leq \gamma^h (U(a) - L(a))\right\}$.

Lemma 11 *This branching factor shrinks as the bounds (L, U) get tighter:*

$$L_2 \leq L_1 \leq V \leq U_1 \leq U_2 \implies \kappa(L_1,U_1) \leq \kappa(L_2,U_2).$$

In particular, $\boldsymbol{\kappa(L,U) \leq \kappa}$.

Theorem 12 *Let $L \leq V \leq U$ monotonic bounds, then planning with L and U in (6) and (7) yields the following simple regret bound:*

$$r_n = \widetilde{\mathcal{O}}\left(n^{-\log\frac{1}{\gamma}/\log \boldsymbol{\kappa(L,U)}}\right).$$

This theorem states that we can potentially improve the performance of the planning algorithm if we manage to find bounds (L, U) that are tighter than the trivial ones at the leaves ∂T_n, which may be possible if we have already seen the states corresponding to this leaves, but it does not explain how to obtain such bounds. In the next subsection, we describe a method to build a sequence of increasingly tight bounds (L_n, U_n), at each planning iteration n. The corresponding regret bound, our main result, is stated in Theorem 16.

4.3. Unrolling the tree to tighten the bounds

In order to reproduce the behaviour of Algorithm 2 on a tree structure, we rely on the following observation: expanding a node s in $\mathcal{G}_n$ simultaneously expands all the paths leading to this node. To account for this observation in the analysis, we will consider an *unrolling* operator T, illustrated in Figure 3, that builds a potentially infinite tree $T(\mathcal{G}_n)$ containing every sequence of action that can be traversed in a graph $\mathcal{G}_n$.

$$T(\mathcal{G}_n) = \{a \in A^h : s_{t+1} \in \mathcal{G}_n \text{ with } s_{t+1} = P(s_t, a_t) \text{ for } 0 \leq t < h\} \tag{9}$$

We analyse GBOP-D though the prism of $T(\mathcal{G}_n)$, which is only used as a theoretical tool.

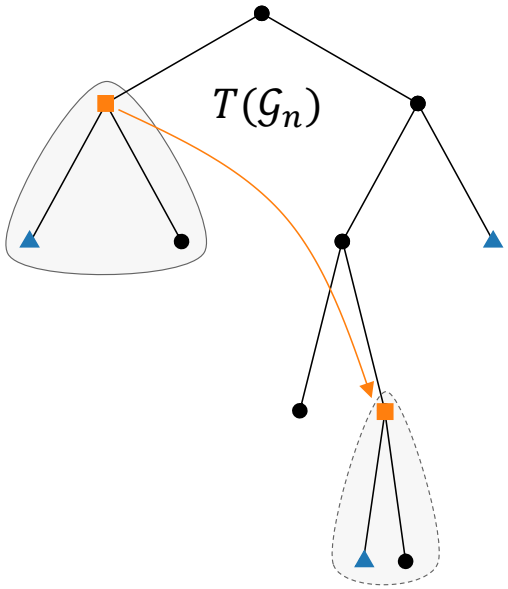


Figure 3: The tree $T(\mathcal{G}_n)$ obtained by unrolling $\mathcal{G}_n$. Contrary to T_n shown in Figure 1, the orange leaf a is expanded at the same time as the internal orange node, which enables to tighten its value bounds $(L_n(a), U_n(a))$ by applying B_n.

We can define the counterpart of the bounds $(\mathcal{L}_n, \mathcal{U}_n)$ in the same way as (4) applied to $T(\mathcal{G}_n)$ rather than T_n, except that the depth d_n of $T(\mathcal{G}_n)$ can now be infinite:

$$L_n = B_n^\infty(0), \qquad U_n = B_n^\infty(V_{\max}). \tag{10}$$

This definition is equivalent to that of GBOP-D in the sense that:

Lemma 13 (Bound equivalence) *For any sequence of action $a \in T(\mathcal{G}_n)$, we have $L_n(a) = \mathcal{L}_n(s(a))$ and $U_n(a) = \mathcal{U}_n(s(a))$.*

In $T(\mathcal{G}_n)$, the unrolling mechanics behave as if any leaf a sharing the same state $s(a)$ as an internal node a' was automatically expanded, and thus had its bound $L_n(a), U_n(a)$ tightened by the Bellman backup B_n to a sub-interval of the trivial bounds $(0, V_{\max})$ that are used in OPD.

The sampling and recommendation rules of GBOP-D amount to running those of OPD on the tree $T(\mathcal{G}_n)$, except that the sampled sequence b_n and recommended sequence a_n can now have infinite depth since $T(\mathcal{G}_n)$ itself can be infinite (we say that a_n and b_n are represented by nodes of infinite depth). In the sequel, we analyse how these rules behave on $T(\mathcal{G}_n)$.

Lemma 14 (Expansion) *Any node a of depth h traversed by the optimistic sampling rule of GBOP-D at iteration n belongs to $T_h^\infty(L_n, U_n)$:*

$$V^\star - V(a) \leq \gamma^h (U_n(a) - L_n(a)). \tag{11}$$

In particular, if the sampling rule samples an infinite sequence $a \in A^\infty$, it is an optimal sequence, and we write that (2) also holds for a with $h = \infty$.

Lemma 15 (Recommendation) *The action a_n recommended by GBOP-D has a simple regret $r_n \leq \frac{\gamma^{d_n}}{1-\gamma}$, where $d_n \in \mathbb{R} \cup \{\infty\}$ is the maximal depth of expanded nodes in $T(\mathcal{G}_n)$.*

Note that even though $T(\mathcal{G}_n)$ can be infinite, there is only one node b_t that is selected for expansion at each iteration $t \leq n$.

4.4. Regret guarantee

In Theorem 12, we assumed that some bounds (L, U) were revealed by an oracle and available from the onset for planning. In (10), we instead built a *sequence* of bounds $(L_n, U_n)_{n\geq 0}$ (10) that is non-increasing in the sense of inclusion, i.e. $0 \leq \cdots \leq L_{n-1} \leq L_n \leq V \leq U_n \leq U_{n-1} \leq \cdots \leq V_{\max}$.

We can consider the sequence $\kappa_n = \kappa(L_n, U_n)$. By Lemma 11, it is non-increasing and lower-bounded by 1, thus converges. Let $\boldsymbol{\kappa_\infty = \lim_{n\to\infty} \kappa(L_n, U_n)} \in [1, K]$.

Theorem 16 *GBOP-D enjoys the following regret bound, with $\kappa_\infty \leq \kappa$:*

$$r_n = \widetilde{\mathcal{O}}\left(n^{-\log \frac{1}{\gamma}/\log \kappa_\infty}\right).$$

Intuitively, κ_∞ should be much lower than κ in problems where trajectories overlap a lot. For instance, it will be the case when two actions cancel each-other out (e.g. moving left or right), or are commutative (e.g. placing pawns on a board game). However, this is merely an intuition. We now show that there exist problem instances in which $\kappa_\infty < \kappa$, which is a legitimate concern since their non-existence would make Theorem 16 trivial.

4.5. Illustrative example

In Proposition 17, we consider a toy MDP $\mathcal{M}$ shown in Figure 4. The transitions are described visually while the rewards are defined as follows: let $0 \leq r^\star \leq \gamma$, and $r^- = r^\star - \frac{\gamma}{1-\gamma} S$, $r^+ = r^\star + S$ with $S = r^\star \left(\frac{1}{\gamma} - 1\right)$. Note that this choice ensures that $r^\star, r^-, r^+$ and S are all in $[0, 1]$.

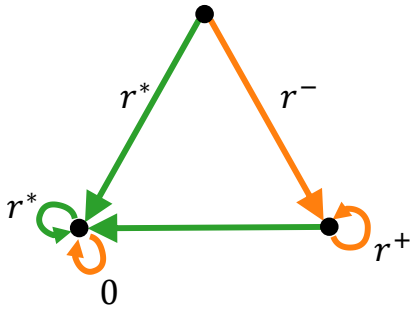


Figure 4: A toy MDP with three states and $K \geq 2$ actions. We start in the top state. The first action a_1 is represented by green arrows, and all other actions $a_2, \ldots, a_K$ are represented by orange arrows. The rewards are shown next to the transitions.

Proposition 17 (Branching factors) *The MDP $\mathcal{M}$ verifies $\boldsymbol{\kappa = K - 1}$ and $\boldsymbol{\kappa_\infty = 1}$.*

This result confirms that Theorem 16 is non-trivial since we exhibit a problem for which $\kappa_\infty < \kappa$ (when $K \geq 3$), and legitimates our attempt to improve planning performances by merging the tree into a graph.

5. Extension to Stochastic Systems

The approach developed in Sections 3 and 4 consists in using state similarity to tighten a pair of lower and upper bounds (L, U) for the value function V. Thus, any planning algorithm that is based on such bounds can benefit from this insight, and any theoretical result based on the validity and rate of convergence of these bounds will be preserved.

Confidence intervals for rewards. When the reward distribution $P(r \mid s, a)$ is stochastic, deviation inequalities can be used to design a confidence interval $[\ell_t(s,a), u_t(s,a)]$ over its expected value $\mathbb{E}[r|s,a]$. For instance, the Chernoff-Hoeffding deviation inequality was used to design confidence intervals in (Kocsis and Szepesvári, 2006; Bubeck and Munos, 2010; Kaufmann and Koolen, 2017). In recent works (Leurent and Maillard, 2019; Jonsson et al., 2020), the tighter Kullback-Leibler confidence interval is preferred:

$$u_t(s,a) \overset{def}{=} \max\left\{v : \mathrm{kl}\big(\hat{r}_t(s,a), v\big) \leq \frac{\beta^r(n_t(s,a), n)}{n_t(s,a)}\right\},$$
$$\ell_t(s,a) \overset{def}{=} \min\left\{v : \mathrm{kl}\big(\hat{r}_t(s,a), v\big) \leq \frac{\beta^r(n_t(s,a), n)}{n_t(s,a)}\right\},$$

where $n_t(s,a)$ is the number of times the transition (s,a) was visited, $\hat{r}_t(s,a)$ is the empirical mean reward, β^r is an exploration function and $\mathrm{kl}(u,v)$ is the binary Kullback-Leibler divergence between Bernoulli distributions: $\mathrm{kl}(u,v) = u \log \frac{u}{v} + (1-u) \log \frac{1-u}{1-v}$.

Confidence region for transitions. Likewise, when the transition distribution $P(s'|s,a)$ is stochastic, a confidence set on the probability vector $p(\cdot|s,a)$ can be defined as $\mathcal{C}_t(s,a) \overset{def}{=} \left\{p \in \Sigma_S : \mathrm{KL}\big(\widehat{p}_t(\cdot|s,a), p\big) \leq \frac{\beta^p(n_t(s,a),n)}{n_t(s,a)}\right\}$, where $\widehat{p}_t(\cdot|s,a) \overset{def}{=} n_t(s,a,\cdot)/n_t(s,a)$ is the empirical distribution, Σ_S is the probability simplex over S, β^p is an exploration function and $\mathrm{KL}(p,q) = \sum_{s \in S} p(s) \log \frac{p(s)}{q(s)}$ is the Kullback-Leibler divergence between categorical distributions.

Bellman operator with stochasticity. In this work, we do not discuss the tuning of β^r, β^p, but simply assume that they are chosen such that the rewards and transitions belong to their confidence regions with sufficiently high probability to obtain performance guarantees for the planning algorithm. For more details on such a choice, refer to (e.g. Leurent and Maillard, 2019; Jonsson et al., 2020). We modify the Definition 2 of $\mathcal{B}_t$ as:

$$\mathcal{B}_t^+(\mathcal{U})(s) = \max_{a \in A}\left[u_t(s,a) + \gamma \max_{q \in \mathcal{C}_t(s,a)} \sum_{s'} q(s'|s,a)\mathcal{U}(s')\right],$$
$$\mathcal{B}_t^-(\mathcal{L})(s) = \max_{a \in A}\left[\ell_t(s,a) + \gamma \min_{q \in \mathcal{C}_t(s,a)} \sum_{s'} q(s'|s,a)\mathcal{L}(s')\right],$$

for all $s \in \mathring{\mathcal{G}}_n$, where the extremums over these KL confidence regions $\mathcal{C}_t(s,a)$ can be computed as explained in (Filippi et al., 2010, Appendix A). Under the event that all confidence regions $[\ell_t(s,a), u_t(s,a)]$ and $\mathcal{C}_t(s,a)$ are valid, the Lemma 5 still holds for $\mathcal{B}_t^-, \mathcal{B}_t^+$.

Structure of the planning algorithm In the deterministic setting, once a transition has been observed, it is known with certainty and doesn't need to be sampled ever again, which is why only external nodes $\partial\mathcal{G}_n$ are sampled in `GBOP-D`. Conversely, in the stochastic setting the expected reward and transition probabilities must be estimated from samples, which implies that internal nodes $\mathring{\mathcal{G}}_n$ must be sampled as well. Then, it is common to adopt and episodic setting where we sample trajectories of a fixed horizon H, tuned depending on the budget n. This is the case in (e.g. Kearns et al., 2002; Kocsis and Szepesvári, 2006; Bubeck and Munos, 2010; Feldman and Domshlak, 2014; Leurent and Maillard, 2019; Jonsson et al., 2020). We also follow this scheme in our proposed `GBOP` algorithm.

Algorithm 3: *Graph-Based Optimistic Planning* (`GBOP`) algorithm.

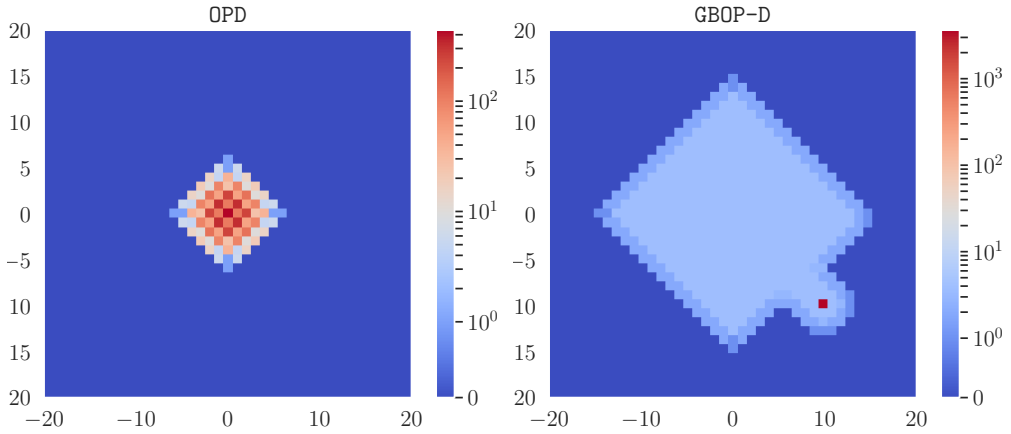
for *trajectory* m *in* $[1, M]$ **do**
  for *time* t *in* $[1, H]$ **do**
    $n \leftarrow (m-1)H + t$.
    Compute the bounds $\mathcal{L}_n = (\mathcal{B}_n^-)^\infty(0)$ and $\mathcal{U}_n = (\mathcal{B}_n^+)^\infty(V_{\max})$.
    $b_t \leftarrow \arg\max_{a \in A} r(s_t, a) + \gamma \mathcal{U}_n(s')$ ▷ `Optimistic sampling rule`
    Simulate $r_t, s_{t+1} \sim P(r, s_{t+1} \mid s_t, b_t)$. Get or create the node s_{t+1} in $\mathcal{G}_{n+1}$, and add an occurrence of the transition (s_t, b_t, r_t, s_{t+1}).
return $\arg\max_{a \in A} r(s,a) + \gamma \mathcal{L}_n(s(a))$. ▷ `Conservative recommendation rule`

6. Numerical Illustration

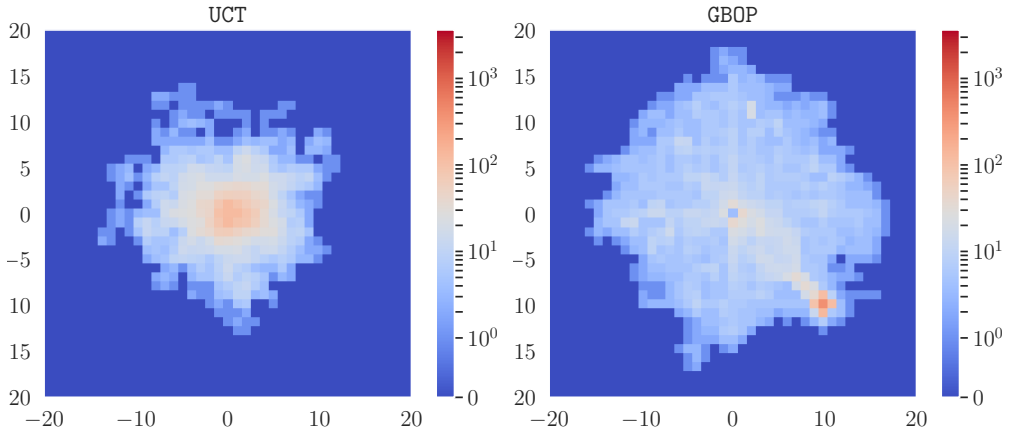
To evaluate the practical benefits of our approach, we compare graph-based and tree-based planning algorithms on two domains.

Gridworld domain. We consider a grid in which the agent starts at $(0,0)$ and can move in $K = 4$ directions. The reward function is 0 everywhere, except in the vicinity of a goal located at $(10,10)$, around which the reward decreases quadratically from 1 to 0 in a ball of radius 5. The Figure 5(a) shows number of times a state is sampled by `OPD` and `GBOP-D`, both run with a budget $n = 5460$ and discount $\gamma = 0.95$. In the absence of rewards, `OPD` samples sequences of actions uniformly (in a breadth-first search manner), which –because of the dynamics structure– results in a non-uniform occupancy of the state space S, where the trajectories concentrate near the starting state. In contrast, `GBOP-D` explores uniformly in S, sampling each state up to four times (from its four neighbours), until it finds the goal vicinity and finally samples the goal location indefinitely. We reproduce the experiment in the stochastic setting by adding noise on the transitions with probability $p = 10\%$, and comparing `GBOP` to `UCT` as we show in Figure 5(b). To quantify these qualitative differences, we define in Figure 5(c) an exploration score: the mean distance $d(s_t, s_0)$ of sampled states to the initial state (exploration) minus the distance $d(s_t, s_g)$ to the goal state (exploitation).

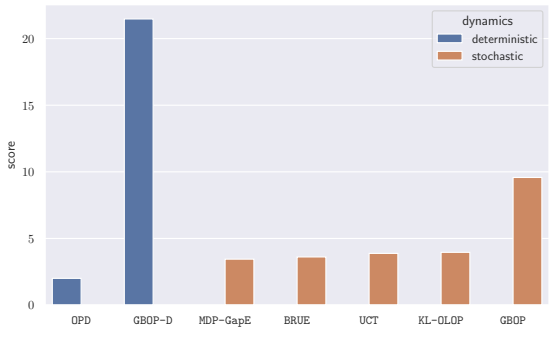
Sailing domain (Vanderbei, 1996). In a second experiment, a boat is sailing in $K = 8$ directions to reach a goal, and suffers a cost (move duration) that depends on the direction



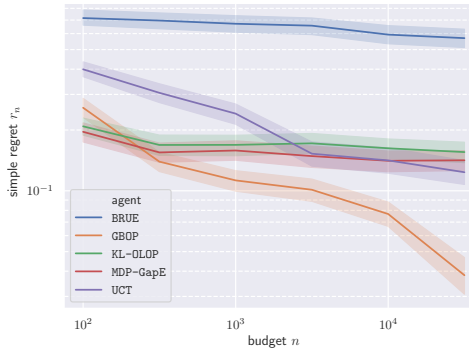
(*a*) State occupancies in a deterministic gridworld.



(*b*) State occupancies in a stochastic gridworld.



(*c*) Exploration scores in the gridworld domains.



(*d*) Simple regret r_n in the sailing domain.

Figure 5: Benchmark of planning performances.

of the wind which follows stochastic dynamics. Figure 5(*d*) shows the evolution of the simple regret r_n of stochastic planning algorithms with respect to the number n of oracle

calls. We compute the mean regret and its 95% confidence interval over 500 simulations. The asymptotic log-log slope σ provides an empirical measurement of the effective branching factor $\kappa_e = \exp(-\log(1/\gamma)/\sigma)$ for each algorithm. We measure that $\sigma \approx -0.04$ and $\kappa_e \approx 3.6$ for `BRUE`, `KL-OLOP`, `MDP-GapE`, `UCT`. In contrast, we measure $\sigma \approx -0.3$ and $\kappa_e \approx 1.2$ for `GBOP`, which suggests that our result of Theorem 16 might generalize to the stochastic setting. Additional experimental results and details are provided in the Supplementary Material.

Acknowledgement

This work was supported by the French Ministry of Higher Education and Research, and CPER Nord-Pas de Calais/FEDER DATA Advanced data science and technologies 2015-2020. We thank Mathieu Seurin and Émilie Kaufmann for their valuable comments.

Conclusion

In this paper, we study a simple yet highly effective variant of the tree-based planning strategies. We prove that leveraging the graph structure induced by states provides a benefit over tree-based algorithms, in the form of an improved regret bound in the deterministic setting, that depends on a smaller difficulty measure. This translates into an enhanced performance in practice, and can be adapted to stochastic planning problems as we show empirically. We believe that revisiting the heart of the MCTS strategy to take into account a graph structure opens exciting novel research directions.

References

Joaquin Ballesteros, Luis Merino, Miguel Angel Trujillo, Antidio Viguria, and Anibal Ollero. Improving the efficiency of online POMDPs by using belief similarity measures. In *Proc. of ICRA*, 2013.

S. Bubeck and R. Munos. Open loop optimistic planning. In *Proc. of COLT*, 2010.

Sébastien Bubeck, Gilles Stoltz, Csaba Szepesvári, and Rémi Munos. Online optimization in x-armed bandits. In *Advances in NIPS*, 2009.

Lucian Busoniu and Rémi Munos. Optimistic planning for markov decision processes. In *Artificial Intelligence and Statistics*, pages 182–189, 2012.

P.-A. Coquelin and R. Munos. Bandit Algorithms for Tree Search. *Proc. of UAI*, 2007.

Rémi Coulom. Efficient Selectivity and Backup Operators in Monte-Carlo Tree Search. In *Proc. of ICCG*, 2006.

Zohar Feldman and Carmel Domshlak. Simple regret optimization in online planning for markov decision processes. *Journal of Artifial Intelligence Research*, 51:165–205, 2014.

S. Filippi, O. Cappé, and A. Garivier. Optimism in Reinforcement Learning and Kullback-Leibler Divergence. In *Allerton Conference on Communication, Control, and Computing*, 2010.

J.-B. Grill, M. Valko, and R. Munos. Blazing the trails before beating the path: Sample-efficient monte-carlo planning. In *Advances in NIPS*, 2016.

Jean-Bastien Grill, Omar Darwiche Domingues, Pierre Ménard, Rémi Munos, and Michal Valko. Planning in entropy-regularized markov decision processes and games. In *Advances in NeurIPS*, 2019.

Jesse Hostetler, Alan Fern, and Tom Dietterich. State aggregation in monte carlo tree search. In *Proc. of AAAI*, 2014.

Jean-Francois Hren and Rémi Munos. Optimistic planning of deterministic systems. In *Proc. of EWRL*, 2008.

Ruitong Huang, Mohammad M. Ajallooeian, Csaba Szepesvári, and Martin Müller. Structured best arm identification with fixed confidence. In *Proc. of ALT*, 2017.

Anders Jonsson, Emilie Kaufmann, Pierre Ménard, Omar Darwiche Domingues, Edouard Leurent, and Michal Valko. Planning in markov decision processes with gap-dependent sample complexity, 2020.

Emilie Kaufmann and Wouter M Koolen. Monte-carlo tree search by best arm identification. In *Advances in NIPS*, pages 4897–4906, 2017.

Michael J. Kearns, Yishay Mansour, and Andrew Y. Ng. A sparse sampling algorithm for near-optimal planning in large markov decision processes. *Machine Learning*, 49(2-3): 193–208, 2002.

Levente Kocsis and Csaba Szepesvári. Bandit based monte-carlo planning. In *Proc. of ECML*, 2006.

Edouard Leurent and Odalric-Ambrym Maillard. Practical open-loop optimistic planning. In *Proc. of ECML-PKDD*, 2019.

R. Munos. *From bandits to Monte-Carlo Tree Search: The optimistic principle applied to optimization and planning.* Foundations and Trends in Machine Learning, 2014.

Rémi Munos. Optimistic optimization of a deterministic function without the knowledge of its smoothness. In *Advances in NIPS*, 2011.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy P. Lillicrap, Karen Simonyan, and Demis Hassabis. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362, 2018.

B. Szorenyi, G. Kedenburg, and R. Munos. Optimistic planning in markov decision processes using a generative model. In *Advances in NIPS*, 2014.

Robert Vanderbei. Optimal sailing strategies, statistics and operations research program. https://vanderbei.princeton.edu/sail/sail.html, 1996.

Supplementary Material

Outline In Appendix A, we provide a proof for every novel result introduced in this paper. Appendix B gives additional insights about the implementation of GBOP-D and GBOP, namely an approximation that guarantees finite-time termination and associated time-complexity. Finally, Appendix C provides additional details about our experiments.

Appendix A. Proofs

A.1. Proof of Lemma 4

Proof The tightening property is directly obtained by definition of monotonicity. Let us show the preservation of monotonicity. Let U a monotonic upper-bound, $a \in A^h$. Then, for any $b \in A$:

$$U(ab) \geq B(U)(ab) \implies r(ab) + \gamma U(ab) \geq r(ab) + \gamma B(U)(ab).$$

Thus, my taking the max on b, $B(U)(a) \geq B^2(U)(a)$. The same can be obtained for a lower-bound L.

The finite time convergence can be obtained by recursion from the leaves to the root, by noticing that if the value of a set of siblings aA is invariant by B, then the value of their parent a is invariant by B^2. ■

A.2. Proof of Lemma 5

Proof The proof of tightening and monotonicity preservation is the same as that of Lemma 4. The contraction property is standard for the Bellman Operator, see e.g. Puterman M., Markov Decision Processes: Discrete Stochastic Dynamic Programming (2005). ■

A.3. Proof of Theorem 9

We recall the main steps of the proof of Hren and Munos (2008).

1. The recommendation a_n has a maximal depth d_n in the tree, and its gap $r_n = V^\star - V(a_{n,1})$ is bounded by $r_n \leq \frac{\gamma^{d_n}}{1-\gamma}$. We need to relate d_n to n.
2. Each expanded node belongs to $T^\infty = \bigcup_{h \geq 0} T_h^\infty$, where

$$T_h^\infty = \left\{ a \in A^h : V^\star - V(a) \leq \frac{\gamma^h}{1-\gamma} \right\}.$$

   Introduce the difficulty measure κ such that $|T_h^\infty| = \mathcal{O}(\kappa^h)$ (the smallest).

3. In the worst case, expanded nodes fully fill the depths of T^∞ up to d_n: $n = \sum_{d=1}^{d_n} n_d \leq C \sum_{d=1}^{d_n} \kappa^d = \begin{cases} \mathcal{O}(d_n) & \text{if } \kappa = 1 \\ \mathcal{O}(\kappa^{d_n}) & \text{else.} \end{cases}$
   Hence $r_n = \begin{cases} \mathcal{O}(\gamma^n) & \text{if } \kappa = 1 \\ \mathcal{O}(\gamma^{\frac{\log n}{\log \kappa}}) = \mathcal{O}(n^{-\frac{\log 1/\gamma}{\log \kappa}}) & \text{else.} \end{cases}$

A.4. Proof of Lemma 11

Proof Let $L_2 \leq L_1 \leq V \leq U_1 \leq U_2$, then $T_h^\infty(L_1, U_1) \subset T_h^\infty(L_2, U_2)$, which implies $|T_h^\infty(L_1, U_1)|^{1/h} \leq |T_h^\infty(L_2, U_2)|^{1/h}$ and the claimed result in the limit $h \to \infty$. ■

A.5. Proof of Theorem 12

In this proof, we temporarily assume that $U = B(U)$ and $L = B(L)$. We follow the same steps as in the proof of the regret of `OPD`.

Remark 18 *It no longer holds that a_n must be of maximal depth d_n. This is due to the fact the exploration bonus $\gamma^h U(a)$ is not depth-wise constant: consider two nodes a, b at the same depth with $R(a) > R(b)$. In `OPD`, both get the same bonus $\gamma^h/(1-\gamma)$, and the node a is expanded first. But with the local bonus, b could be expanded in priority rather than a, if its own bonus is sufficiently higher than that of a, precisely if $R(a) + \gamma^h U(a) < R(b) + \gamma^h U(b)$. For instance, $U(a) = 0$ when a is known to be a terminal state while b can lead to future rewards. If after expanding and exploring the subtree of b we find out that $V(b) = 0$, we still return the recommendation a, which is of non-maximal depth.*

The regret bound still holds, however. First, notice that:

Lemma 19 (Expansion) *Whenever a node a of depth h is expanded by the optimistic algorithm, its first action a_1 enjoys a simple regret $V(a^\star) - V(a_1) \leq \gamma^h (U(a) - L(a))$.*

Proof Let t be the time of expansion of a, it holds that $\overline{U}_t(b) \leq \overline{U}_t(a)$ for all $b \in \partial T_t$, in particular those in a branch starting by an optimal action $a^\star$. Since $U = B(U)$ and $L = B(L)$, we also have $\overline{U}_t(a^\star) = \max_{b \in a^\star A^*} \overline{U}_t(b) \leq \overline{U}_t(a)$, and $\overline{L}_t(a_1) = \max b \in a_1 A^* \overline{L}_t(b) \geq \overline{L}_t(a)$. Thus, $V(a^\star) - V(a_1) \leq \overline{U}_t(a^\star) - \overline{L}_t(a_1) \leq \overline{U}_t(a) - \overline{L}_t(a) = \gamma^h (U(a) - L(a))$. ■

Lemma 20 (Recommendation) *The recommended action a_n has a simple regret $r_n \leq \frac{\gamma^{d_n}}{1-\gamma}$, where d_n is the maximal depth of T_n.*

Proof Let i a node of maximal depth d_n, and consider the recommended node a_n at time n, of depth d. In particular, $\overline{L}_n(a_n) \geq \overline{L}_n(i)$, and since $(\overline{L}_t)_t$ is non-decreasing we also have $\overline{L}_n(i) \geq \overline{L}_t(i)$. At the time t when i is expanded, we have $\overline{U}_t(a_n) \leq \overline{U}_t(i)$, and since $(\overline{U}_t)_t$ is non-increasing we also have $\overline{U}_n(a_n) \leq \overline{U}_t(a_n)$. We can conclude with Lemma 19 applied to a_n: $r_n \leq \gamma^d (U(a_n) - L(a_n) = \overline{U}_n(a_n) - \overline{L}_n(a_n) \leq \overline{U}_t(a_n) - \overline{L}_n(i) \leq \overline{U}_t(i) - \overline{L}_t(i) = \gamma^{d_n} (U(i) - L(i)$, which yields the claimed bound since $U(i) - L(i) \leq V_{\max} - 0$. ■

Lemma 21 (Near-optimal nodes) *Every node expanded by (6) is in* $T^\infty(L,U) = \bigcup_{h\geq 0} T_h^\infty(L,U)$.

Proof Let a be a node of depth h expanded at round n, then $\overline{U}_n(a) \geq \overline{U}_n(b)$ for all $b \in \partial T_n$. Thus, since $U = B(U)$, we have $\overline{U}(a) = \overline{B(U)}(\emptyset) = B(U)(s_0) \geq V(s_0) = V^\star$. Thus, $V^\star - V(a) \leq \overline{U}(a) - \overline{L}(a) = \gamma^h(U(a) - L(a))$. ■

Finally, we can move on to the proof of Theorem 12. Let n_d be the number of expanded nodes of depth d, by Lemma 21 we have $n_d \leq |T_d^\infty(L,U)| \leq C\kappa(L,U)^d$. Thus,

$$n = \sum_{d=1}^{d_n} n_d \leq C \sum_{d=0}^{d_n} \kappa(L,U)^d = C\frac{\kappa(L,U)^{d_n+1} - 1}{\kappa(L,U) - 1}$$

Hence, $d_n \geq C' \frac{\log n}{\log \kappa(L,U)}$, which along with Lemma 19 gives the claimed bound.

Note that if L, U are monotonic bounds that do not verify $L = B(L)$ and $U = B(U)$, then planning with $B(L), B(U)$ instead will yield the proved bound with a branching factor $\kappa(B(L), B(U))$, and since $L \leq B(L) \leq V \leq B(U) \leq U$ we have $\kappa(B(L), B(U)) \leq \kappa(L,U)$, which still gives

$$r_n = \mathcal{O}\left(n^{-\frac{\log 1/\gamma}{\log \kappa(L,U)}}\right);$$

A.6. Proof of Lemma 13

Proof

We first show that if U is equivalent to $\mathcal{U}$, meaning that for any sequence $a \in T(\mathcal{G}_n)$ we have $U(a) = \mathcal{U}(s(a))$, then $B_n(U)$ is equivalent to $\mathcal{B}_n(\mathcal{U})$.

By definition of $T(\mathcal{G}_n)$, any sequence of action $a \in T(\mathcal{G}_n)$ corresponds to a path $s_0, a_0, \ldots, s_{h-1}, a_{h-1}, s_h$ in $\mathcal{G}_n$. If $a \in \partial T(\mathcal{G}_n)$, then necessarily $s(a) \in \partial\mathcal{G}_n$, and both are unchanged by B_n and $\mathcal{B}_n$ respectively. Conversely, if $a \in \mathring{T}(\mathcal{G}_n)$, then $s(a) \in \mathring{\mathcal{G}}_n$ by construction. Thus, $B_n(U)(a) = \max_{b\in A} r(s(a),b) + \gamma U(ab) = \max_{b\in A} r(s(a),b) + \gamma\mathcal{U}(s(ab))$ (by hypothesis) $= \max_{b\in A} r(s(a),b) + \gamma\mathcal{U}(P(s(a),b)) = \mathcal{B}_n(\mathcal{U}_n)(s(a))$.

By induction, for any $k > 0$ $B_n^k(U)$ is equivalent to $\mathcal{B}_n^k(\mathcal{U})$, and at the limit $k \to \infty$ it comes that U_n is equivalent to $\mathcal{U}_n$. The same result can be shown similary for L_n and $\mathcal{L}_n$. ■

A.7. Proof of Lemma 14

We start by showing a preliminary lemma.

Lemma 22 (Bounds of sequence values) *The bounds* $(\overline{L}_n, \overline{U}_n)$ *on the value of sequences of actions verify are respectively non-decreasing and non-increasing with respect to* n*, and verify: for all* $a \in A^*$, $\overline{U_n}(a) = \max_{a' \in aA^\infty} \overline{U}(a')$.

Proof The second property can be easily shown by induction using the fact that U_n and L_n are fixed-points of B_n by definition. Applying this equation at each depth h gives the result. From this observation, we can deduce that $\overline{L}_n$ is increasing with n. Indeed, since when

$T(\mathcal{G}_n)$ is expanded with additional nodes compared to $T(\mathcal{G}_{n-1})$, the leaves a of $T(\mathcal{G}_{n-1})$ with previous value $L_{n-1}(a) = 0$ are updated to $L_n(a) = \max_b r(s(a), b) \geq 0 = L_{n-1}(a)$, and this increase at the leaves is then propagated through $\max_{a' \in aA^\infty}$ to any internal node a. Thus, L_n is non-decreasing and likewise, U_n is non-increasing with respect to n. The same is obtained directly of the bounds on sequence values $(\overline{L}_n, \overline{U}_n)$. ■

Which enables us to proceed to the proof of Lemma 14.

Proof Let t be the time of expansion of a, it holds that $\overline{U}_t(b) \leq \overline{U}_t(a)$ for all $b \in T(\mathcal{G}_n)$. In particular for b in a branch starting by an optimal action $a^\star$ $\overline{U}_t(a) \geq \max_{b \in a^\star A^*} \overline{U}_t(b) = \overline{U}_t(a^\star)$. Thus, $V(a^\star) - V(a) \leq \overline{U}_t(a^\star) - \overline{L}_t(a) \leq \overline{U}_t(a) - \overline{L}_t(a) = \gamma^h(U_t(a) - L_t(a))$. ■

A.8. Proof of Lemma 15

Proof Let i an expanded node of maximal depth $d_n \in \mathbb{R} \cup \{\infty\}$, and consider the recommended node a_n at time n, of depth $d \in \mathbb{R} \cup \{\infty\}$. In particular, $\overline{L}_n(a_n) \geq \overline{L}_n(i)$, and since $(\overline{L}_t)_t$ is non-decreasing we also have $\overline{L}_n(i) \geq \overline{L}_t(i)$. At the time t when i is expanded, we have $\overline{U}_t(a_n) \leq \overline{U}_t(i)$, and since $(\overline{U}_t)_t$ is non-increasing we also have $\overline{U}_n(a_n) \leq \overline{U}_t(a_n)$. We can conclude with Lemma 14 applied to a_n: $r_n \leq V^\star - V(a_n) \leq \gamma^d(U(a_n) - L(a_n) = \overline{U}_n(a_n) - \overline{L}_n(a_n) \leq \overline{U}_t(a_n) - \overline{L}_n(i) \leq \overline{U}_t(i) - \overline{L}_t(i) = \gamma^{d_n}(U_t(i) - L_t(i)$, which yields the claimed bound since $U(i) - L(i) \leq V_{\max} - 0$. ■

A.9. Proof of Theorem 16

Proof Let $\kappa' > \kappa_\infty$. Since $\kappa(L_n, U_n) \to \kappa_\infty$, there exists $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$, $\kappa(L_n, U_n) \leq \kappa'$. By Lemma 14, at each iteration n the expanded node must belong to $T^\infty(L_n, U_n)$. Let $n \geq n_0$, and define $d_0 = \min\{d \in \mathbb{N} : \exists t \in [n_0, n], b_t \in A^d\}$. By definition, for all $d \geq d_0$, any expanded node of depth d was expanded at a time $t \geq n_0$, and thus $b_t \in T_t^\infty \subset T_{n_0}^\infty$. We denote n_d the number of expanded nodes of depth d. If $d_n = \infty$,then $r_n = 0$ and the bound holds. Else, we obtain

$$n = \sum_{d=0}^{d_0-1} n_d + \sum_{d=d_0}^{d_n} n_d \leq C_0 + C_1 \sum_{d=d_0}^{d_n} (\kappa')^d \leq C_0 + C_1'(\kappa')^{d_n}$$

And since $r_n \leq \frac{\gamma^{d_n}}{1-\gamma}$ by Lemma 15, we obtain the claimed bound.

Moreover, given a history of observed transitions up to iteration n, the bounds U_n, L_n obtained from (10) on the unrolled tree $T(\mathcal{G}_n)$ are tighter than those of (4) since $T_n \subset T(\mathcal{G}_n)$, which implies by Lemma 11 that $\kappa(L_n, U_n) \leq \kappa$. We obtain $\kappa_\infty \leq \kappa$ at the limit. ■

A.10. Proof of Proposition 17

The Figure 6 shows the planning tree corresponding to the MDP $\mathcal{M}$. Whenever the action a_1 is taken (in green) the resulting subtree is represented by a leaf node $s^\star$ of value $V^\star = \frac{r^\star}{1-\gamma}$.

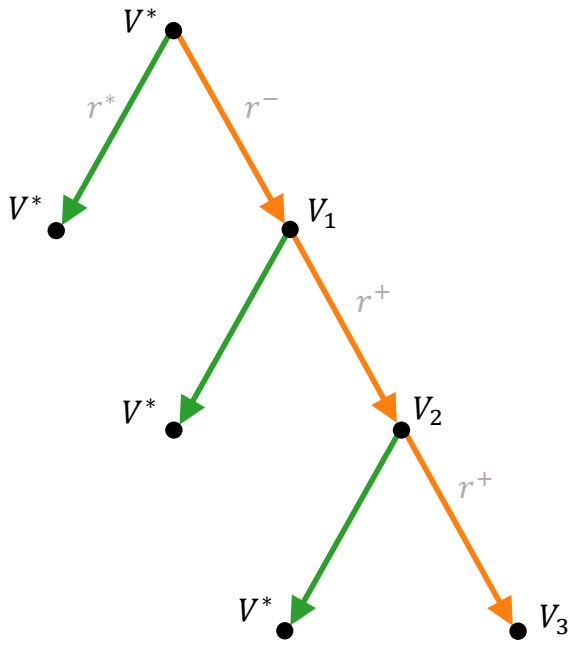


Figure 6: Planning tree of the MDP $\mathcal{M}$ of Figure 4

.

When, in contrast, we take a sequence of actions among $a_2 \dots a_K$ (in red), we stay in the state s^+ and denote V_h the corresponding value at depth h.

Lemma 23 *Any sequence of actions in $A \setminus a_1$ is in T^∞.*

Proof Any such sequence of actions yields the sequence of rewards $r^-, r^+, \dots, r^+$. and end up in the state s^+ with value at least $V^\star$ (obtained by further taking a_1 indefinitely). Thus its value V_h verifies,

$$\begin{aligned} V_h &\geq \sum_{t=0}^{h-1} \gamma^t r_t + \gamma^h V^\star \\ &= r^- - r^+ + \sum_{t=0}^{h-1} \gamma^t r^+ + \gamma^h V^\star \\ &= (-\frac{\gamma}{1-\gamma} - 1)S + \frac{1-\gamma^h}{1-\gamma}(r^\star + S) + \gamma^h V^\star \\ &= V^\star - S\frac{\gamma^h}{1-\gamma} \geq V^\star - \frac{\gamma^h}{1-\gamma} \end{aligned}$$

■

We can directly conclude that $\kappa \geq \limsup |\{a_2, \dots, a_K\}^h|^{1/h} = K - 1$.

Now, consider the nodes expanded by GBOP-D. The first expansion is that of the root, which discovers $s^\star$ and s^+. In the absence of information on these two state, the bound

$V_{\max}$ is used and the first action a_1 gets a higher $\overline{U}$ that any other action $a_2, \dots, a_K$ since $r^\star \geq r^-$. Hence, at the second iteration, the node a_1 gets expanded. At this point, the self-loop of the state $s^\star$ is discovered, which means that form now on the bounds verify $L_n(a_1) = V^\star = U_n(a_1)$ for $n \geq 2$, which means that $L_n(a_1 A^*) - U_n(a_1 A^*) = 0$. The nodes $a_2, \dots, a_K$ can be expanded at most once before the entire MDP is discovered and $L_n = V = U_n$ over the entire tree, which means that T_n^∞ is the set of optimal nodes, i.e. the nodes in the only optimal sequence $a_1^\star$. Hence, $\kappa_\infty = 1$.

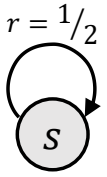
Appendix B. Implementation Details

In this section, we provide more details about the implementation of GBOP-D and GBOP. First, we discuss how two procedures can be approximated so that they terminate in finite time, and study the impact of this approximation on the regret guarantees. Second, we propose a lazy implementation of the bounds computation through $\mathcal{B}_n^\infty$ that only considers a subset of nodes to update.

B.1. Termination

B.1.1. Bounds computation

The bounds computation step $\mathcal{B}_n^\infty$ (line 1 of GBOP-D) can converge in infinite time whenever $\mathcal{G}_n$ contains a loop, as shown in Figure 7. We consider the effect of stopping early after a fixed number of iterations $k(\varepsilon, \gamma)$.



k	0	1	$\cdots$	k
$\mathcal{U} = \mathcal{B}^k(V_{\max})(s)$	$V_{\max}$	$\frac{1}{2} + \gamma V_{\max}$		$\frac{1}{2}(1-\gamma^k)V_{\max} + \gamma^k V_{\max}$
$\mathcal{L} = \mathcal{B}^k(0)(s)$	0	$\frac{1}{2}$		$\frac{1}{2}(1-\gamma^k)V_{\max}$

Figure 7: **Top**: a simple looping MDP with $|S| = |A| = 1$ after having observed a single transition ($n = 1$). **Bottom**: the sequence of bounds $\mathcal{B}_1^k(0)$ and $\mathcal{B}_1^k(V_{\max})$. They converge geometrically to their limit $\mathcal{U}_1 = \mathcal{L}_1 = V = \frac{1}{2}V_{\max}$, thus in infinite time.

Proposition 24 (Time complexity of bounds computation) *An ε-approximation of $(\mathcal{L}_n, \mathcal{U}_n)$ can be computed by applying $\mathcal{B}_n$ for a finite number $k(\varepsilon, \gamma)$ of iterations, with*

$$k(\varepsilon, \gamma) = \log_\gamma \frac{1}{\varepsilon(1-\gamma)}.$$

Proof $\mathcal{B}_n$ is a γ-contraction by Lemma 5, and $\mathcal{U}_n$ (resp $\mathcal{L}_n$) is at a distance (in $\|\dot{\|}_\infty$) at most $V_{\max}$ of the initial value bound $V_{\max}$ (resp 0). Thus, the k^{th} application of $\mathcal{B}_n$

decreases this error by a factor γ^k, which gives the result. ■

The impact of using an ε-approximation of $(\mathcal{L}_n, \mathcal{U}_n)$ during planning is the following:

Proposition 25 (Effect of early stopping) *Denote the approximate bounds $(\hat{\mathcal{L}}_n, \hat{\mathcal{U}}_n)$ obtained by applying $\mathcal{B}_n^{k(\varepsilon,\gamma)}$ instead of $\mathcal{B}_n^{\infty}$, and likewise $(\hat{L}_n, \hat{U}_n)$ in their tree version obtained by applying $B_n^{k(\varepsilon,\gamma)}$ instead of B_n^{∞}. Then, running GBOP-D with $\hat{\mathcal{L}}_n, \hat{\mathcal{U}}_n$ gives the following regret:*

$$r_n = \widetilde{\mathcal{O}}\left(n^{-\log\frac{1}{\gamma}/\log \hat{\kappa}_\infty}\right),$$

with

$$\kappa_\infty \leq \hat{\kappa}_\infty \overset{def}{=} \lim_{n\to\infty} \kappa(\hat{L}_n, \hat{U}_n) \leq \kappa.$$

Moreover, the approximation gap $\kappa_\infty - \hat{\kappa}_\infty$ is non-increasing with respect to ε.

It is difficult to control more explicitly the gap between κ_∞ and $\hat{\kappa}_\infty$, which might be discontinuous with ε.

Proof Note that $\hat{L}_n$ and $\hat{U}_n$ are valid monotonic bounds on V, verifying

$$0 \leq \hat{L}_n \leq L_n \leq V \leq U_n \leq \hat{U}_n \leq V_{\max}.$$

Thus, Lemma 22 holds with the difference that we only have an inequality $\overline{U_n}(a) \geq \max_{a' \in aA^\infty} \overline{U}(a')$ rather than an equality, by monotonicity but non-invariance by B_n. However, this was the actual inequality used in Lemma 14, which still holds by replacing L_n, U_n by their approximation $\hat{L}_n, \hat{U}_n$. Likewise, Lemma 15 holds. The proof of Theorem 16, can be written with the modification that expanded nodes belong to $T_h^\infty(\hat{L}_n, \hat{U}_n))$, which gives the claimed bound.

As ε decreases, $k(\varepsilon, \gamma)$ increases, which means by Lemma 4 that $(\hat{\mathcal{L}}_n, \hat{\mathcal{U}}_n)$ get tighter and $\hat{\kappa}_\infty$ shrinks by Lemma 11. It reaches its minimum κ_∞ when $\varepsilon = 0$. ■

Thus, we observe that there is a *trade-off* between the time complexity $k(\varepsilon)$ and the sample complexity $\hat{\kappa}_\infty$: decreasing one increases the other. We illustrate this trade-off empirically in Figure 8.

Note that OPD uses d_n iterations of B_n, which corresponds to a tuning of ε with n: $\varepsilon_n = \frac{\gamma^{d_n}}{1-\gamma} = \mathcal{O}(n^{-\frac{\log 1/\gamma}{\log \kappa}})$.

B.1.2. Sampling rule

The sampling rule of GBOP-D (line 2 of GBOP-D) can yield an infinite sequence b_n. We propose to stop the sampling after a fixed depth d_n^+.

Proposition 26 (Time complexity of sampling) *Consider the variant of GBOP-D where we stop the sampling rule when reaching a fixed depth d_n^+ chosen polynomial with n:*

$$d_n^+ = \lceil \alpha n^\beta \rceil, \text{ with } \alpha, \beta > 0$$

Then, the regret bound of Theorem 16 (or that of Proposition 25 when using early stopping in the bounds computation) still holds.

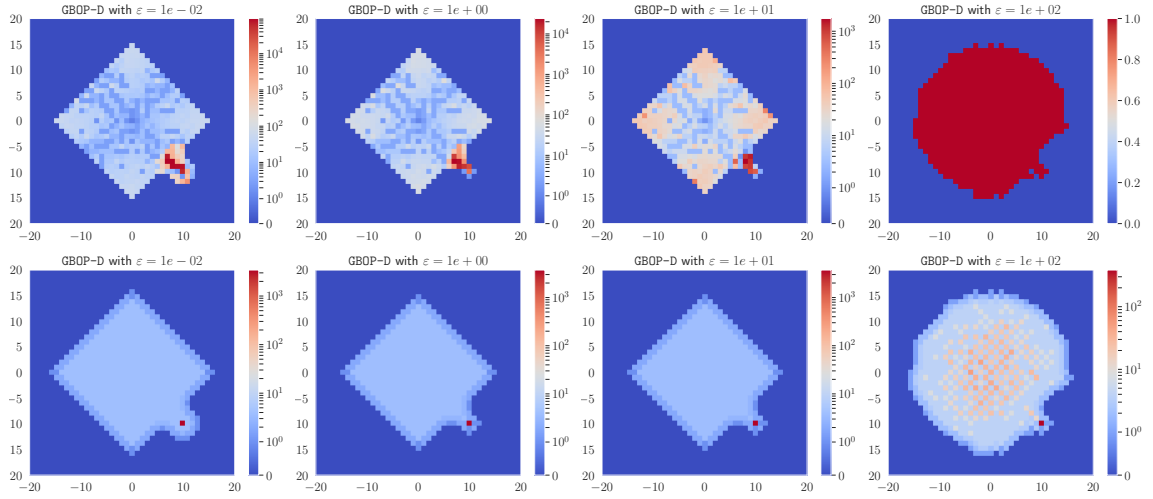


Figure 8: **Top row**: Map showing the number of updates triggered by $\mathcal{B}$ at each expanded state for various values of ε. **Bottom row**: Map showing the corresponding state occupations. The rightmost column corresponds to $\varepsilon \geq V_{\max} = 20$, *i.e.* every approximation stops after a single iteration.

Note that this is not too constraining compared to OPD, for which the sampling rule complexity d_n is upper-bounded by n. Hence, by choosing $\alpha = \beta = 1$, GBOP-D preserve the same complexity as OPD in the worst case.

Proof Let $\kappa' > \boldsymbol{\kappa_\infty}$ (or $\kappa' > \hat{\boldsymbol{\kappa}}_\infty$ under approximate bounds). In the proof of Theorem 16, it is shown that the maximum depth d_n of an expanded node is at least $d_n^- \overset{def}{=} \log_{\kappa'} \frac{n - C_0}{C_1'}$, which allows to conclude with Lemma 15 that $r_n = \mathcal{O}(\gamma^{d_n}) = \mathcal{O}(\gamma^{d_n^-})$. By choosing d_n^+ polynomial, we have that d_n^+ is greater than d_n^- for n sufficiently high. Thus, by stopping the sampling after reaching a depth d_n^+, we have that $r_n \leq \gamma^{\min\{d_n, d_n^+\}}/(1-\gamma) = \mathcal{O}(\gamma^{d_n^-}) = \mathcal{O}(n^{\frac{-\log 1/\gamma}{\log \kappa'}})$ ■

B.2. Efficient implementation of $\mathcal{B}_n^\infty$

The bounds $\mathcal{L}_n$ and $\mathcal{U}_n$ are computed by fixed-point iteration of $\mathcal{B}_n$ from the trivial bounds $(0, V_{\max})$. The naive implementation of $\mathcal{B}_n$ requires to iterate over the whole set of state-action pairs in $\mathcal{G}_n$. Two ideas can be used to increase the efficiency of both steps:

(i) Instead of starting the iteration with the trivial bounds, the previous estimate $\mathcal{L}_{n-1}, \mathcal{U}_{n-1}$ can be used instead at iteration n. Since these bounds are closer to their limit ($0 \leq \mathcal{L}_{n-1} \leq \mathcal{L}_n$ and $\mathcal{U}_n \leq \mathcal{U}_{n-1} \leq V_{\max}$), the fixed-point iteration will converge quicker.

(ii) In particular, since $\mathcal{L}_{n-1}$ and $\mathcal{U}_{n-1}$ are invariant by $\mathcal{B}_n$, the the only nodes modified by a supplementary application of $\mathcal{B}_n$ are the parents of only updated node: the expanded

state s_n. Once its value is updated by $\mathcal{B}_n$, the same reasoning can be applied for the next iteration of $\mathcal{B}_n$: only its predecessors can be updated. Thus, we can keep track of a set q of states that can be updated, for every application of $\mathcal{B}_n$.

These idea are formalised in Algorithm 4. Note that the criterion $\|\mathcal{U}_n^{k+1} - \mathcal{U}_n^k\| \leq \frac{1-\gamma}{\gamma}\varepsilon$ is used to detect that the limit $\mathcal{U}_n^\infty$ is approximated with accuracy ε, and stems from $\mathcal{B}_n$ being a γ-contraction:

Proof $\|\mathcal{U}_n^k - \mathcal{U}_n^\infty\| \leq \gamma\|\mathcal{U}_n^{k+1} - \mathcal{U}_n^\infty\| \leq \gamma\|\mathcal{U}_n^{k+1} - \mathcal{U}_n^k\| + \gamma\|\mathcal{U}_n^k - \mathcal{U}_n^\infty\|$, with $\|\mathcal{U}_n^{k+1} - \mathcal{U}_n^k\| \leq \frac{1-\gamma}{\gamma}\varepsilon$, thus $\|\mathcal{U}_n^k - \mathcal{U}_n^\infty\| \leq \varepsilon$. ■

Algorithm 4: A queue-based implementation of $\mathcal{B}_n^\infty$.
Input: Initial bound $\mathcal{U}_{n-1}$, expanded node s_n, accuracy ε
Output: An ε-approximation of $\mathcal{U}_n$

```
U_n ← U_{n-1}
q ← [s_n]
while q is not empty do
    s' ← Pop the first node from the queue q
    U' ← B_n(U_n)(s')                                   ▷ Node backup
    if U' − U_n > ((1−γ)/γ)ε then                       ▷ Stopping rule
        Push the predecessors s of s' to the queue q    ▷ Propagation rule
    end
    U_n(s') ← U'
end
return U_n
```

Appendix C. Experimental Details

In this section, we provide additional details and results about the experiments discussed in Section 6.

Parameters. In every experiment, we used $\gamma = 0.95$. In GBOP-D and GBOP, the fixed accuracy $\varepsilon = 1 \times 10^{-2}$ was used for computing $\mathcal{B}_n^\infty$, and the sampling ruled was stopped after reaching the depth $d_n^+ = n$ (see Appendix B). Regarding the tuning of confidence intervals in GBOP, since the rewards are deterministic and the transitions are stochastic in both domains we used $\beta^r(n_t(s,a),n) = 0$ and $\beta^p(n_t(s,a),n) = \log n$, following the recommendations of (Leurent and Maillard, 2019; Jonsson et al., 2020). The maximal size B of the support of the transitions $P(s'|s,a)$ was also used ($B = 4$ in the gridworld domain and $B = 3$ in the sailing domain) to accelerate the computations of the confidence region $\mathcal{C}_t$ for transitions, as explained in (Jonsson et al., 2020).

Additional results. We show in Figure 9 some occupancy plots for additional planning algorithms in the stochastic Gridworld domain. In Figure 10, we compare the trees T_n expanded by OPD and KL-OLOP to the unrolled graph $T(\mathcal{G}_n)$ of GBOP-D in the deterministic gridworld domain. For clarity of the figure, we only display the nodes selected for expansion and not the entire $T(\mathcal{G}_n)$ as in Figure 3, since it is infinite and fractal. We observe that OPD

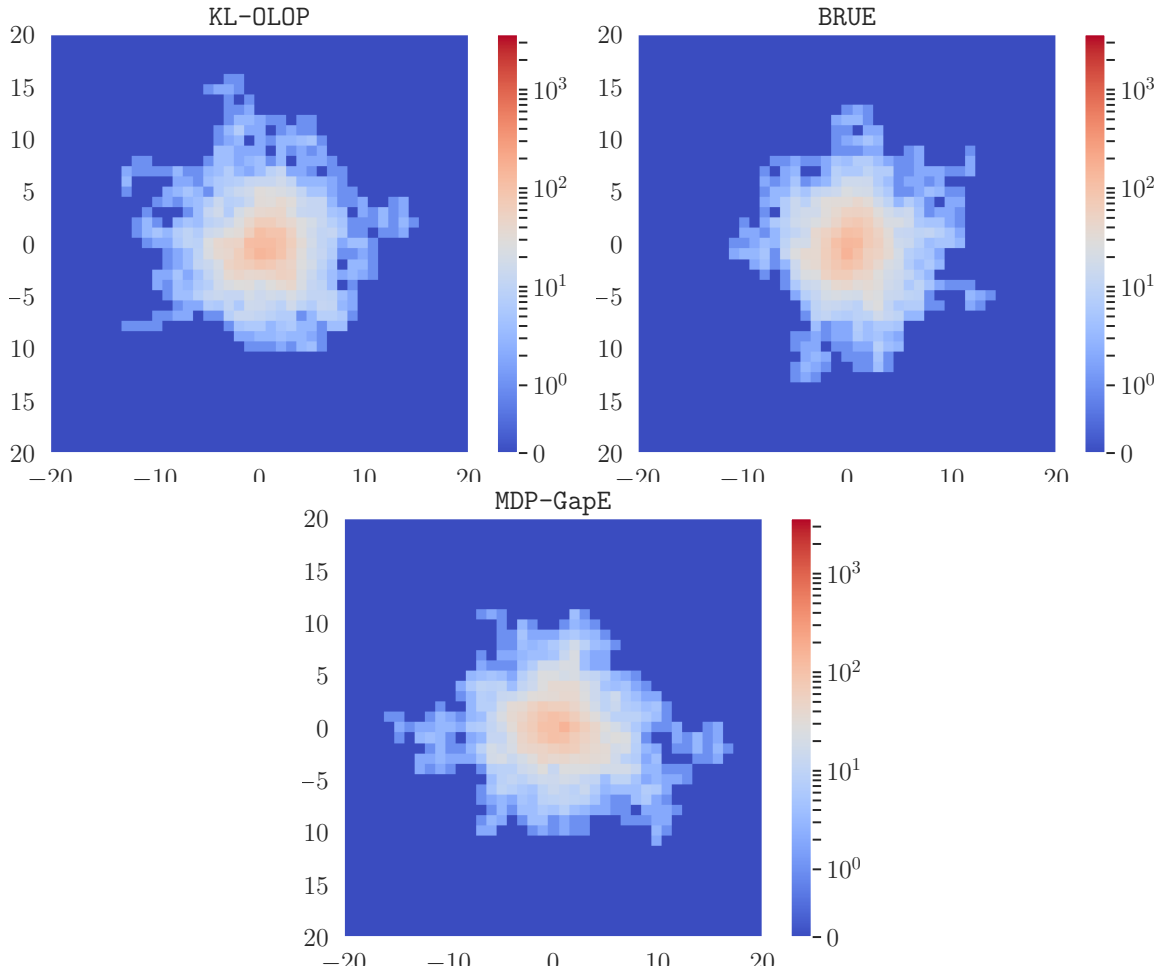


Figure 9: State occupancies of other planning algorithms in a stochastic gridworld.

explores uniformly in the space of sequences, which results in a concentrated exploration in the state space as seen in Figure 5(*a*). This phenomenon is similar to the concentration properties of martingales. KL-OLOP behaves similarly, but allocates its budget n in fewer trajectories of higher length than OPD, which results in a sparser tree and slightly better exploration (compare Figure 5(*a*) to Figure 9). In contrast, GBOP-D expands a very sparse and unbalanced tree $T(\mathcal{G}_n)$ which corresponds to uniform exploration in the state space, and allows to explore deeper for the same budget n (The tree is only shown up to depth 12, but continues much deeper since the optimal transition is sampled indefinitely once it is discovered). In particular, the paths towards the goal are sampled many times, while other algorithms are still balanced at the root in terms of number of visits.

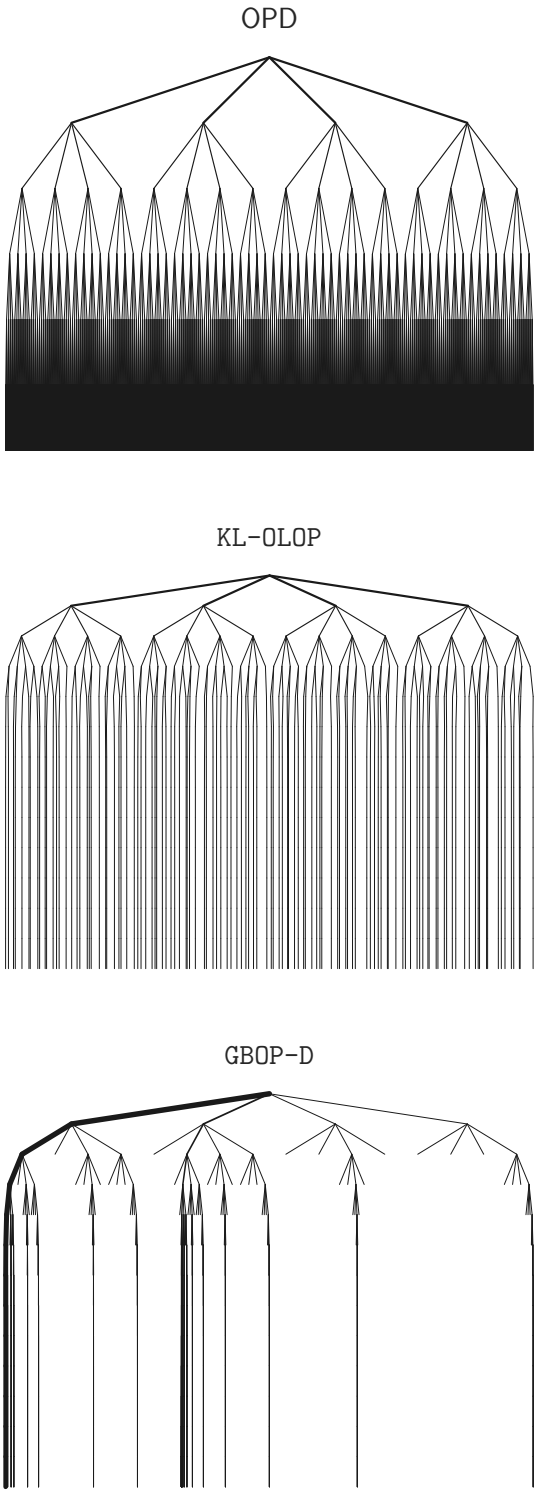


Figure 10: Trees expanded by OPD, by KL-OLOP, and sequences of actions sampled by GBOP-D. The width of edges is proportional to the number of visits.