



HAL
open science

Blazing the trails before beating the path: Sample-efficient Monte-Carlo planning

Jean-Bastien Grill, Michal Valko, Remi Munos

► **To cite this version:**

Jean-Bastien Grill, Michal Valko, Remi Munos. Blazing the trails before beating the path: Sample-efficient Monte-Carlo planning. *Neural Information Processing Systems*, Dec 2016, Barcelona, Spain. hal-01389107v1

HAL Id: hal-01389107

<https://inria.hal.science/hal-01389107v1>

Submitted on 28 Oct 2016 (v1), last revised 2 Feb 2017 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Blazing the trails before beating the path: Sample-efficient Monte-Carlo planning

Jean-Bastien Grill

SequeL team, INRIA Lille - Nord Europe, France

jean-bastien.grill@inria.fr

Michal Valko

michal.valko@inria.fr

Rémi Munos

Google DeepMind, UK*

munos@google.com

Abstract

You are a robot and you live in a Markov decision process (MDP) with a finite or an infinite number of transitions from state-action to next states. You got brains and so you *plan* before you act. Luckily, your roboparents equipped you with a generative model to do some *Monte-Carlo planning*. The world is waiting for you and you have no time to waste. You want your planning to be efficient. *Sample-efficient*. Indeed, you want to exploit the possible structure of the MDP by exploring only a subset of states reachable by following near-optimal policies. You want guarantees on sample complexity that depend on a measure of the quantity of near-optimal states. You want something, that is an extension of Monte-Carlo sampling (for estimating an expectation) to problems that alternate maximization (over actions) and expectation (over next states). But you do not want to StOP with exponential running time, you want something simple to implement and computationally efficient. You want it all and you want it now. You want **TrailBlazer**.

1 Introduction

We consider the problem of sampling-based planning in a Markov decision process (MDP) when a *generative model* (oracle) is available. This approach, also called Monte-Carlo planning or Monte-Carlo tree search (see e.g., [11]), has been popularized in the game of computer Go [7, 8, 13] and shown impressive performance in many other high dimensional control and game problems [4]. In the present paper, we provide a sample complexity analysis of a new algorithm called **TrailBlazer**.

Our assumption about the MDP is that we possess a generative model which can be called from any state-action pair to generate rewards and transition samples. Since making a call to this generative model has a cost, be it a numerical cost expressed in CPU time (in simulated environments) or a financial cost (in real domains), our goal is to *use this model as parsimoniously as possible*.

Following *dynamic programming* [2], planning can be reduced to an approximation of the (optimal) value function, defined as the maximum of the expected sum of discounted rewards: $\mathbb{E} \left[\sum_{t \geq 0} \gamma^t r_t \right]$, where $\gamma \in [0, 1)$ is a known *discount factor*. Indeed, if an ε -optimal approximation of the value function at any state-action pair is available, then the policy corresponding to selecting in each state the action with the highest approximated value will be $\mathcal{O}(\varepsilon / (1 - \gamma))$ -optimal [3].

Consequently, in this paper, we focus on the a near-optimal approximation of the value function *for a single given state* (or state-action pair). In order to assess the performance of our algorithm we measure its *sample complexity* defined as the number of oracle calls, given that we guarantee its *consistency*, i.e., that with probability at least $1 - \delta$, **TrailBlazer** returns an ε -approximation of the value function as required by the probably approximately correct (PAC) framework.

*on the leave from SequeL team, INRIA Lille - Nord Europe, France

We use a *tree representation* to represent the set of states that are reachable from any initial state. This tree alternates maximum (**MAX**) nodes (corresponding to actions) and average (**AVG**) nodes (corresponding to the random transition to next states). We assume the number K of actions is finite. However, the number N of possible next states is either *finite* or *infinite* (which may be the case when the state space is infinite), and we will report results in both the finite N and the infinite case. The root node of this planning tree represents the current state (or a state-action) of the MDP and its value is the maximum (over all policies defined at **MAX** nodes) of the corresponding expected sum of discounted rewards. Notice that by using a tree representation, we do not use the property that some state of the MDP can be reached by different paths (sequences of states-actions). Therefore, this state will be represented by different nodes in the tree. We could potentially merge such duplicates to form a graph instead. However, for simplicity, we choose not to merge these duplicates and keep a tree, which could make the planning problem harder. To sum up, our goal is to return, with probability $1 - \delta$, an ε -accurate value of the root node of this planning tree while using as low number of calls to the oracle as possible. Our contribution is an algorithm called **TrailBlazer** whose sampling strategy depends on the specific structure of the MDP and for which we provide *sample complexity* bounds in terms of a new *problem-dependent measure of the quantity of near-optimal nodes*. Before describing our contribution in more detail we first relate our setting to what has been around.

1.1 Related work

Kocsis and Szepesvári [11] introduced the UCT algorithm (upper-confidence bounds for trees). UCT is efficient in computer Go [7, 8, 13] and a number of other control and game problems [4]. UCT is based on generating trajectories by selecting in each **MAX** node the action that has the highest upper-confidence bound (computed according to the UCB algorithm of Auer et al. [1]). UCT converges asymptotically to the optimal solution, but its sample complexity can be worst than doubly-exponential in $(1/\varepsilon)$ for some MDPs [12]. One reason for this is that the algorithm can expand very deeply the apparently best branches but may lack sufficient exploration, especially when a narrow optimal path is hidden in a suboptimal branch. As a result, this approach works well in some problems with a specific structure but may be much worse than a uniform sampling in other problems.

On the other hand, a uniform planning approach is safe for all problems. Kearns et al. [10] generate a sparse look-ahead tree based on expanding all **MAX** nodes and sampling a finite number of children from **AVG** nodes up to a fixed depth that depends on the desired accuracy ε . Their sample complexity is² of the order of $(1/\varepsilon)^{\log(1/\varepsilon)}$, which is *non-polynomial* in $1/\varepsilon$. This bound is better than that for UCT in a worst-case sense. However, as their look-ahead tree is built in a *uniform* and *non-adaptive* way, this algorithm fails to benefit from a potentially favorable structure of the MDP.

An improved version of this sparse-sampling algorithm by Walsh et al. [15] cuts sub-optimal branches in an adaptive way but unfortunately does not come with an improved bound and stays non-polynomial even in the simple Monte Carlo setting for which $K = 1$.

Although the sample complexity is certainly non-polynomial in the worst case, it *can be polynomial* in some specific problems. First, for the case of finite N , the sample complexity is polynomial and Szörényi et al. [14] show that a uniform sampling algorithm has complexity at most $(1/\varepsilon)^{2+\log(KN)/(\log(1/\gamma))}$. Notice that the product KN represents the *branching factor* of the look-ahead planning tree. This bound could be improved for problems with specific reward structure or transition smoothness. In order to do this, we need to design non-uniform, *adaptive* algorithm that captures the possible structure of the MDP when available, while making sure that in the worst case, we do not perform worse than a uniform sampling algorithm.

The case of deterministic dynamics ($N = 1$) and rewards considered by Hren and Munos [9] has a complexity of order $(1/\varepsilon)^{(\log \kappa)/(\log(1/\gamma))}$, where $\kappa \in [1, K]$ is the branching factor of the subset of near-optimal nodes.³ The case of stochastic rewards has been considered by Bubeck and Munos [5] but with the difference that the goal was not to approximate the optimal value function but the value of the best *open-loop* policy which consists in a sequence of actions independent of states. Their sample complexity is $(1/\varepsilon)^{\max(2, (\log \kappa)/(\log 1/\gamma))}$.

²neglecting exponential dependence in γ

³nodes that need to be considered in order to return a near-optimal approximation of the value at the root

In the case of general MDPs, Buşoniu and Munos [6] consider the case of a fully known model of the MDP. For any state-action, the model returns the expected reward and the set of all next states (assuming N is finite) with their corresponding transition probabilities. In that case, the complexity is $(1/\varepsilon)^{\log \kappa / (\log(1/\gamma))}$, where $\kappa \in [0, KN]$ can again be interpreted as a branching factor of the subset of near-optimal nodes. These approaches use the *optimism in the face of uncertainty* principle whose applications to planning have been set in stone by Munos [12]. **TrailBlazer** is different. It is *not optimistic* by design: To avoid voracious demand for samples it does not balance the upper-confidence bounds of all possible actions. This is crucial for polynomial sample complexity in the infinite case.

The work that is most related to ours is StOP by Szörényi et al. [14] which considers the planning problem in MDPs with a generative model. Their complexity bound is of the order of $(1/\varepsilon)^{2 + \log \kappa / (\log(1/\gamma)) + o(1)}$, where $\kappa \in [0, KN]$ is a problem-dependent quantity. However, their κ defined as $\lim_{\varepsilon \rightarrow 0} \max(\kappa_1, \kappa_2)$ (in their Theorem 2) is somehow difficult to interpret as a measure of the quantity of *near-optimal nodes*. Moreover, StOP is not computationally efficient as it requires to identify the *optimistic policy* which requires computing an upper bound on the value of *any* possible policy, whose number is exponential in the number of **MAX** nodes, which itself is exponential in the planning horizon. Although they suggest (in their Appendix F) a computational improvement, this version is not analyzed. Finally, unlike in the present paper, StOP does not consider the case $N = \infty$ of an unbounded number of states.

1.2 Our contribution

Our main result is **TrailBlazer**, an algorithm with a bound on the number of samples required to return a high-probability ε -approximation of the root node whether the number of next states N is finite or infinite. The bounds use a problem-dependent quantity (κ or b) that measures the quantity of near-optimal nodes. We now summarize the results.

Finite number of next states ($N < \infty$): The sample complexity of **TrailBlazer** is of the order of⁴ $(1/\varepsilon)^{\max(2, \log(N\kappa) / \log(1/\gamma)) + o(1)}$, where $\kappa \in [1, K]$ is related to the branching factor of the set of near-optimal nodes (precisely defined later).

Infinite number of next states ($N = \infty$): The complexity of **TrailBlazer** is $(1/\varepsilon)^2$, if $b > 2$ where b parametrise a bound on the probability distribution of children gaps at **AVG** nodes. Notice that b can be greater than 2 even if the planning problem is very challenging.⁵ We also state our contributions in specific settings in comparison to previous work.

- For the case $N < \infty$, we improve over the best-known previous worst-case bound with an exponent (to $1/\varepsilon$) of $\max(2, \log(NK) / \log(1/\gamma))$ instead of $2 + \log(NK) / \log(1/\gamma)$ reported by Szörényi et al. [14].
- For the case $N = \infty$, we identify properties of the MDP (when $b > 2$) under which the sample complexity is of order (in $1/\varepsilon^2$). This is as good as Monte-Carlo sampling and for this reason **TrailBlazer is a natural extension of Monte-Carlo sampling** (where all nodes are **AVG**) to stochastic control problems (where **MAX** and **AVG** nodes alternate). Also, no previous algorithm reported a polynomial bound in the case of $N = \infty$.
- In highly structured MDPs where there are non-vanishing action-gaps⁶ from any state along near-optimal policies, then b is unbounded and the complexity is $\tilde{O}(1/\varepsilon^2)$.
- In MDPs with deterministic transitions ($N = 1$) but stochastic rewards our bound is $(1/\varepsilon)^{\max(2, \log \kappa / (\log 1/\gamma))}$ which is similar to the bound achieved by Bubeck and Munos [5] in a similar setting (open-loop policies).
- In the evaluation case without control ($K = 1$) our bound provides a complexity of $1/\varepsilon^2$, which is as good as Monte-Carlo sampling. Notice that this bound holds even when $N = \infty$.
- Finally **TrailBlazer** is easy to implement and is numerically efficient.

⁴neglecting logarithmic terms in ε and δ

⁵since when $N = \infty$ the actual branching factor of the set of reachable nodes is infinite

⁶defined as the difference in values of best and second-best actions

2 Monte-Carlo planning with a generative model

Setup We operate on a planning tree \mathcal{T} . Each node of \mathcal{T} from the root down is alternatively either an average (AVG) or a maximum (MAX) node. For any node s , $\mathcal{C}[s]$ is the set of its children. We consider trees \mathcal{T} for which the cardinality of $\mathcal{C}[s]$ for any MAX node s is bounded by K . The cardinality N of $\mathcal{C}[s]$ for any AVG node s can be either finite, $N < \infty$, or infinite. We consider *both* cases. **TrailBlazer** applies to both situations. We provide performance guarantees for a general case and possibly tighter, N -dependent guarantees in the case of $N < \infty$. We assume that we have a generative model of the transitions and rewards: Each AVG node s is associated with a transition, a random variable $\tau_s \in \mathcal{C}[s]$ and a reward, a random variable $r_s \in [0, 1]$.

Objective For any node s , we define the value function $\mathcal{V}[s]$ as the optimum over policies π (giving a successor to all MAX nodes) of the sum of discounted expected rewards playing policy π ,

$$\mathcal{V}[s] = \sup_{\pi} \mathbb{E} \left[\sum_{t \geq 0} \gamma^t r_{s_t} \mid s_0 = s, \pi \right],$$

where $\gamma \in (0, 1)$ is the *discount factor*. If s is an AVG node, \mathcal{V} satisfies the following Bellman equation,

$$\mathcal{V}[s] = \mathbb{E}[r_s] + \gamma \sum_{s' \in \mathcal{C}[s]} p(s'|s) \mathcal{V}[s'].$$

If s is a MAX node, then $\mathcal{V}[s] = \max_{s' \in \mathcal{C}[s]} \mathcal{V}[s']$.

The planner has access to the oracle which can be called for any AVG node s to either get a reward r or a transition τ which are two independent random variables identically distributed as r_s and τ_s respectively.

With the notation above, **our goal is to estimate the value $\mathcal{V}[s_0]$ of the root node s_0 using the smallest possible number of oracle calls.** More precisely, given any δ and ε , we want to output a value $\mu_{\varepsilon, \delta}$ such that $\mathbb{P}[|\mu_{\varepsilon, \delta} - \mathcal{V}[s_0]| > \varepsilon] \leq \delta$ using the smallest possible number of oracle calls $n_{\varepsilon, \delta}$. The number of calls is the *sample complexity* of the algorithm.

2.1 Blazing the trails with TrailBlazer

To fulfill the above objective, our **TrailBlazer** constructs a planning tree \mathcal{T} which is, at any time, a finite subset of the potentially infinite tree. Only the already visited nodes are in \mathcal{T} and explicitly represented in memory. Taking the object-oriented paradigm, each node of \mathcal{T} is a persistent object with its own memory which can receive and perform calls respectively from and to other nodes. A node can potentially be called several times (with different parameters) during the run of **TrailBlazer** and may reuse (some of) its stored (transition and reward) samples. In particular, after node s receives a call from its parent, node s may perform internal computation by calling its own children in order to return a real value to its parent.

Pseudocode of **TrailBlazer** is in Figure 1 along with the subroutines for MAX nodes in Figure 3 and AVG nodes in Figure 2. A node (MAX or AVG) is called with two parameters m and ε , which represent some requested properties of the returned value: m controls the desired *variance* and ε the desired *maximum bias*. We now describe the MAX and AVG node subroutines.

```

1: Input:  $\delta, \varepsilon$ 
2: Set:  $\eta \leftarrow \gamma^{1/\max(2, \log(1/\varepsilon))}$ 
3: Set:  $\lambda \leftarrow 2 \log(\varepsilon(1-\gamma))^2 \frac{\log(\frac{\log(K)}{1-\eta})}{\log(\gamma/\eta)}$ 
4: Set:  $m \leftarrow (\log(1/\delta) + \lambda)/((1-\gamma)^2 \varepsilon^2)$ 
5: Use:  $\delta$  and  $\eta$  as global parameters
6: Output:
    $\mu \leftarrow$  call the root with parameters  $(m, \varepsilon/2)$ 

```

Figure 1: **TrailBlazer**

```

1: Input:  $m, \varepsilon$ 
2: Initialization: {Only executed on first call}
3: SampledNodes  $\leftarrow \emptyset$ ,
4:  $r \leftarrow 0$ 
5: Run:
6: if  $\varepsilon \geq 1/(1-\gamma)$  then
7:   Output: 0
8: end if
9: if |SampledNodes| >  $m$  then
10:   ActiveNodes  $\leftarrow$  SampledNodes(1 :  $m$ )
11: else
12:   while |SampledNodes| <  $m$  do
13:      $\tau \leftarrow$  {new sample of next state}
14:     SampledNodes.append( $\tau$ )
15:      $r \leftarrow r +$ {new sample of reward}
16:   end while
17:   ActiveNodes  $\leftarrow$  SampledNodes
18: end if {At this point, |ActiveNodes| =  $m$ }
19: for all unique nodes  $s \in$  ActiveNodes do
20:    $k \leftarrow$  #occurrences of  $s$  in ActiveNodes
21:    $\nu \leftarrow$  call  $s$  with parameters  $(k, \varepsilon/\gamma)$ 
22:    $\mu \leftarrow \mu + \nu k/m$ 
23: end for
24: Output:  $\gamma \mu + r/|\text{SampledNodes}|$ 

```

Figure 2: **AVG node**

MAX nodes A **MAX** node s keeps a lower and an upper bound of its children values which with high probability simultaneously hold at all times. It sequentially calls its children with different parameters in order to get more and more precise estimates of their values. Whenever the upper bound of one child becomes lower than the maximum lower bound, this child is discarded. This process can stop in two ways: 1) The set \mathcal{L} of the remaining children shrunk enough such that there is a single child b^* left. In this case, s calls b^* with the same parameters that s received and uses the output of b^* as its own output. 2) The precision we have on the value of the remaining children is high enough. In this case, s returns the highest estimate of the children in \mathcal{L} . Note that the **MAX** node is eliminating actions to identify the best. Any other best-arm identification algorithm for bandits can be adapted instead.

```

1: Input:  $m, \varepsilon$ 
2:  $\mathcal{L} \leftarrow$  all children of the node
3:  $\ell \leftarrow 1$ 
4: while  $|\mathcal{L}| > 1$  and  $U \geq (1 - \eta)\varepsilon$  do
5:    $U \leftarrow \frac{2}{1-\gamma} \sqrt{\frac{\log(K\ell/(\delta\varepsilon)) + \gamma/(\eta-\gamma) + \lambda + 1}{\ell}}$ 
6:   for  $b \in \mathcal{L}$  do
7:      $\mu_b \leftarrow$  call  $b$  with  $(\ell, U\eta/(1-\eta))$ 
8:   end for
9:    $\mathcal{L} \leftarrow \left\{ b: \mu_b + \frac{2U}{1-\eta} \geq \sup_j \left[ \mu_j - \frac{2U}{1-\eta} \right] \right\}$ 
10:   $\ell \leftarrow \ell + 1$ 
11: end while
12: if  $|\mathcal{L}| > 1$  then
13:   Output:  $\mu \leftarrow \max_{b \in \mathcal{L}} \mu_b$ 
14: else  $\{ \mathcal{L} = \{b^*\} \}$ 
15:    $b^* \leftarrow \arg \max_{b \in \mathcal{L}} \mu_b$ 
16:    $\mu \leftarrow$  call  $b^*$  with  $(m, \eta\varepsilon)$ 
17:   Output:  $\mu$ 
18: end if

```

Figure 3: **MAX** node

AVG nodes Every **AVG** node s keeps a list of all the children that it already sampled and a reward estimate $r \in \mathbb{R}$. Note that the list may contain the same child multiple times (this is particularly true for $N < \infty$). After receiving a call with parameters (m, ε) , s checks if $\varepsilon \geq 1/(1-\gamma)$. If this condition is verified, then it returns zero. If not, s considers the first m sampled children and potentially samples more children from the generative model if needed. For every child s' in this list, s calls it with parameters $(k, \varepsilon/\gamma)$, where k is the number of times a transition toward this child was sampled. It returns $r + \gamma\mu$, where μ is the average of all the children estimates.

Anytime algorithm **TrailBlazer** is *naturally anytime*. It can be called with slowly decreasing ε , such that m is always increased only by 1, without having to throw away **any** previously collected samples. Executing **TrailBlazer** with ε' and then with $\varepsilon < \varepsilon'$ leads to the same amount of computation as immediately running **TrailBlazer** with ε .

3 Cogs whirring behind

Before diving into the analysis we explain the ideas behind **TrailBlazer** and the choices made.

Tree-based algorithm The number of policies the planner can consider is *exponential* in the number of states. This leads to two major challenges. First, reducing the problem to multi-arm bandits on the set of the policies would hurt. When a reward is collected from a state, all the policies which could reach that state are affected. Therefore, it is useful to share the information between the policies. The second challenge is computational as it is infeasible to keep all policies in memory.

These two problems immediately vanish with just how **TrailBlazer** is formulated. Contrary to Szörényi et al. [14], we do not represent the policies explicitly or update them simultaneously to share the information, but we store all the information directly in the planning tree we construct. Indeed, by having all the nodes being separate entities that store their own information, we can share information between policies without explicitly having to enforce it.

We steel ourselves for the detailed understanding with the following two arguments. They shed light from two different angles on the very same key point: Do not refine more paths than you need to!

Delicate treatment of uncertainty First, we give intuition about the *two parameters* which measure the requested *precision* of a call. The output estimate μ of any call with parameters (m, ε) verifies the following property (conditioned on a high-probability event),

$$\forall \lambda \quad \mathbb{E} \left[e^{\lambda(\mu - \mathcal{V}[s])} \right] \leq \exp \left(\alpha + \varepsilon |\lambda| + \frac{\sigma^2 \lambda^2}{2} \right), \quad \text{with } \sigma^2 = O(1/m) \text{ and constant } \alpha. \quad (1)$$

This awfully looks like the definition of μ being uncentered sub-Gaussian, except that instead of λ in the exponential function, there is $|\lambda|$ and there is a λ -independent constant α . Inequality 1 implies that the absolute value of the bias of the output estimate μ is bounded by ε ,

$$|\mathbb{E}[\mu] - \mathcal{V}[s]| \leq \varepsilon.$$

As in the sub-Gaussian case, the second term $\frac{1}{2}\sigma^2\lambda^2$ is a variance term. Therefore, ε controls the maximum bias of μ and $1/m$ control its sub-variance. In some cases, getting high-variance or low-variance estimate matters less as it is going to be averaged later with other independent estimates by an ancestor **AVG** node. In this case we prefer to query for high variance rather than a low one, in order to decrease sample complexity.

From σ and ε it is possible to deduce a confidence bounds on $|\mu - \mathcal{V}[s]|$ by typically summing the bias ε and a term proportional to the standard deviation $\sigma = O(1/\sqrt{m})$. Previous approaches [14, 5] consider a *single* parameter, representing the width of this high-probability confidence interval. **TrailBlazer** is different. In **TrailBlazer**, the nodes can perform high-variance and low-bias queries but can also query for both low-variance and low-bias. **TrailBlazer** treats these two types of queries *differently*. This is the whetstone of **TrailBlazer** and the reason why it is not optimistic.

Refining few paths In this part we explain the condition $|\text{SampledNodes}| > m$ in Figure 2, which is crucial for our approach and results. First notice, that as long as **TrailBlazer** encounters only **AVG** nodes, it behaves just like Monte-Carlo sampling — without the **MAX** nodes we would be just doing a simple averaging of trajectories. However, when **TrailBlazer** encounters a **MAX** node it *locally* uses more samples around this **MAX** node, *temporally* moving away from a Monte-Carlo behavior. This enables **TrailBlazer** to compute the best action at this **MAX** node. Nevertheless, once this best action is identified with high probability, the algorithm should behave again like Monte-Carlo sampling. Therefore, **TrailBlazer** *forgets* the additional nodes, sampled just because of the **MAX** node, and only keeps in memory the first m ones. This is done with the following line in Figure 2,

$$\text{ActiveNodes} \leftarrow \text{SampledNodes}(1 : m).$$

Again, while additional transitions were useful for some **MAX** node parents to decide which action to pick, they are discarded once this choice is made. Note that they can become useful again if an ancestor becomes unsure about which action to pick and needs more precision to make a choice. This is an important difference between **TrailBlazer** and some previous approaches like UCT where all the already sampled transitions are equally refined. This treatment enables us to provide *polynomial* bounds on the sample complexity for some special cases even in the infinite case ($N = \infty$).

4 **TrailBlazer** is good and cheap — consistency and sample complexity

In this section, we start by our consistency result, stating that **TrailBlazer** outputs a correct value in a PAC (probably approximately correct) sense. Later, we define a measure of the problem difficulty which we use to state our sample-complexity results. We remark that the following consistency result holds whether the state space is finite or infinite.

Theorem 1. *For all ε and δ , the output $\mu_{\varepsilon, \delta}$ of **TrailBlazer** called on the root s_0 with (ε, δ) verifies*

$$\mathbb{P}[|\mu_{\varepsilon, \delta} - \mathcal{V}[s_0]| > \varepsilon] < \delta.$$

4.1 Definition of the problem difficulty

We now define a measure of problem difficulty that we use to provide our sample complexity guarantees. We define a set of *near-optimal* nodes such that exploring only this set is enough to compute an optimal policy. Let s' be a **MAX** node of tree \mathcal{T} . For any of its descendants s , let $c_{\rightarrow s}(s') \in \mathcal{C}[s']$ be the child of s' in the path between s' and s . For any **MAX** node s , we define

$$\Delta_{\rightarrow s}(s') = \max_{x \in \mathcal{C}[s']} V(x) - V(c_{\rightarrow s}(s')).$$

$\Delta_{\rightarrow s}(s')$ is the difference of the sum of discounted rewards stating from s' between an agent playing optimally and one playing first the action toward s and then optimally.

Definition 1 (near-optimality). We say that a node s of depth h is near-optimal, if for all $h' < h$

$$\Delta_{\rightarrow s}(s_{h'}) \leq 12 \frac{\gamma^{h-h'}}{1-\gamma}$$

with $s_{h'}$ the ancestor of s of depth h' . Let \mathcal{N}_h be the set of all near-optimal nodes of depth h .

Remark 1. Notice that the subset of near-optimal nodes contains all required information to deduce the value of the root. And in the case $N = \infty$ when $p(s|s') = 0$ for all s, s' then our definition of near-optimality nodes leads to the smallest subset such that the knowledge of the rewards on this set only is sufficient to deduce the value of the root.

We prove that with probability $1 - \delta$, **TrailBlazer** only explore near-optimal nodes. Thus the size of the subset of near-optimal nodes directly reflects the sample complexity of **TrailBlazer**.

In Appendix C, we discuss the negatives of other potential definitions of near-optimal nodes.

4.2 Sample complexity in the finite case

We first state our result where the set of the **AVG** children nodes is *finite* and bounded by N .

Definition 2. We define $\kappa \in [1, K]$ as the smallest number such that

$$\exists C \forall h, \quad |\mathcal{N}_h| \leq C(N\kappa)^h.$$

Notice that since the total number of nodes of depth h is bounded by $(KN)^h$, κ is upper-bounded by K , the maximum number of **MAX**'s children. However κ can be as low as 1 in cases when the set of near-optimal nodes is small. An example of such case is when the gap at **MAX** nodes is uniformly lower-bounded by some $\Delta > 0$ (see more discussion below).

Theorem 2. There exists $C > 0$ and K such that for all $\varepsilon > 0$ and $\delta > 0$, with probability $1 - \delta$, the sample-complexity of **TrailBlazer** (the number of calls to the generative model before the algorithm terminates) is

$$n(\varepsilon, \delta) \leq C(1/\varepsilon)^{\max(2, \frac{\log(N\kappa)}{\log(1/\gamma)} + o(1))} (\log(1/\delta) + \log(1/\varepsilon))^\alpha,$$

where $\alpha = 5$ when $\log(N\kappa)/\log(1/\gamma) \geq 2$ and $\alpha = 3$ otherwise.

This provides a problem-dependent sample-complexity bound, which already in the worst case ($\kappa = K$) improves over the best-known worst-case bound $\tilde{O}\left((1/\varepsilon)^{2 + \frac{\log(KN)}{\log(1/\gamma)}}\right)$ [14].

4.3 Sample complexity in the infinite case

Since the previous bound depends on N , it does not apply to the infinite case with $N = \infty$. We now provide a sample complexity result in the case $N = \infty$. However, notice that when N is bounded, then *both* results apply.

Assumption 1. There exists $a > 0$ and $b > 2$ such that for any node s ,

$$\mathbb{P}[\Delta_s \leq x] \leq ax^b,$$

with Δ_s the gap at a **MAX** node s defined as

$$\Delta_s = \min_{i \in \mathcal{C}[s], i \neq i^*} \mathcal{V}[i^*] - \mathcal{V}[i] \text{ and } i^* = \arg \max_{i \in \mathcal{C}[s]} \mathcal{V}[i].$$

Theorem 3. Under Assumption 1, there exists $C > 0$ such that for all $\varepsilon > 0$ and $\delta > 0$ the expected sample complexity of **TrailBlazer** verifies

$$\mathbb{E}[n(\varepsilon, \delta)] \leq C \frac{(\log(1/\delta) + \log(1/\varepsilon))^3}{\varepsilon}.$$

Note that this bound holds in expectation, contrary to the previous high-probability result.

Remark 2. In the case where there exists a lower bound $\Delta \geq \Delta_{\min}$ on the gap at **MAX** nodes then $\mathbb{P}[\Delta_s \leq x] \leq \frac{x^3}{\Delta_{\min}^3}$ and we recover a $(1/\varepsilon^2)$ rate.

Acknowledgements The research presented in this paper was supported by French Ministry of Higher Education and Research, Nord-Pas-de-Calais Regional Council, a doctoral grant of École Normale Supérieure in Paris, Inria and Carnegie Mellon University associated-team project EduBand, and French National Research Agency projects ExTra-Learn (n.ANR-14-CE24-0010-01) and BoB.

References

- [1] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.
- [2] Richard Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ, 1957.
- [3] Dimitri Bertsekas and John Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA, 1996.
- [4] Cameron B. Browne, Edward Powley, Daniel Whitehouse, Simon M. Lucas, Peter I. Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez, Spyridon Samothrakis, and Simon Colton. A survey of Monte Carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in Games*, 4(1):1–43, 2012.
- [5] Sébastien Bubeck and Rémi Munos. Open-loop optimistic planning. In *Conference on Learning Theory*, 2010.
- [6] Lucian Buşoniu and Rémi Munos. Optimistic planning for Markov decision processes. In *International Conference on Artificial Intelligence and Statistics*, 2012.
- [7] Rémi Coulom. Efficient selectivity and backup operators in Monte-Carlo tree search. *Computers and games*, 4630:72–83, 2007.
- [8] Sylvain Gelly, Wang Yizao, Rémi Munos, and Olivier Teytaud. Modification of UCT with patterns in Monte-Carlo Go. Technical report, Inria, 2006. URL <https://hal.inria.fr/inria-00117266>.
- [9] Jean-Francois Hren and Rémi Munos. Optimistic Planning of Deterministic Systems. In *European Workshop on Reinforcement Learning*, 2008.
- [10] Michael Kearns, Yishay Mansour, and Andrew Y. Ng. A sparse sampling algorithm for near-optimal planning in large Markov decision processes. In *International Conference on Artificial Intelligence and Statistics*, 1999.
- [11] Levente Kocsis and Csaba Szepesvári. Bandit-based Monte-Carlo planning. In *European Conference on Machine Learning*, 2006.
- [12] Rémi Munos. From bandits to Monte-Carlo tree search: The optimistic principle applied to optimization and planning. *Foundations and Trends in Machine Learning*, 7(1):1–130, 2014.
- [13] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.
- [14] Balázs Szörényi, Gunnar Kedenburg, and Rémi Munos. Optimistic planning in Markov decision processes using a generative model. In *Neural Information Processing Systems*, 2014.
- [15] Thomas J Walsh, Sergiu Goschin, and Michael L Littman. Integrating sample-based planning and model-based reinforcement learning. *AAAI Conference on Artificial Intelligence*, 2010.

A Consistency

First, we comment on the behavior of the nodes when called multiple times.

Remark 3. Notice that the output of the call to a node s with parameters (m, ε) and global δ does not depend on when this call was performed. That means if this call is performed again later in the execution with the same parameters (m, ε) and δ , the output would be the same. This is because, first, the **MAX** nodes do not store anything so that their behavior only depends on s, m, ε , and δ . Second, the **AVG** nodes store the already sampled rewards, transitions and the order they were sampled but when called with parameters (m, ε) and δ , it returns the average of the m first sampled children only.

From now we consider any δ, η , and γ , such that $0 < \gamma < \eta < 1$ and $0 < \delta < 1/2$.

We define a random variable $\mu_s(m, \varepsilon)$ to be the output of the node s called with parameters (m, ε) and global parameters δ and η . Notice that following Remark 3, $\mu_s(m, \varepsilon)$ does depend on $s, m, \varepsilon, \delta, \eta$ and all the transitions and rewards of itself and its descendants but does not depend on the current state of the algorithm.

We define the function $U(m, \varepsilon)$ such that the value $U(m, \varepsilon) + \varepsilon$ is the width of a high-probability confidence interval for $\mu_s(m, \varepsilon)$ which is used in **MAX** node,

$$U(m, \varepsilon) \stackrel{\text{def}}{=} \frac{2}{1 - \gamma} \sqrt{\frac{\log(K\ell/(\delta\varepsilon)) + \gamma/(\eta - \gamma) + 1}{m}}.$$

We define $\ell_{\max}(\varepsilon) \stackrel{\text{def}}{=} \max\{\ell : U(\ell, \varepsilon) \geq (1 - \eta)\varepsilon\}$. By definition of **TrailBlazer**, this is the maximum value that the children of a **MAX** node would be ever called with. We also define $\bar{\ell}(m, \varepsilon) \stackrel{\text{def}}{=} \max(m, \ell_{\max}(\varepsilon))$

We now construct event \mathcal{B}_s for which Lemma 1 proves that it holds with high probability and that on \mathcal{B}_s , **TrailBlazer** works well.

Definition 3. For any node s , $\varepsilon > 0$, and $m \in \mathbb{N}^+$,

- If $\varepsilon \geq 1/(1 - \gamma)$,

$$\mathcal{B}_s(m, \varepsilon) \stackrel{\text{def}}{=} \Omega.$$

- Else, if s is a **MAX** node,

$$\mathcal{B}_s(m, \varepsilon) \stackrel{\text{def}}{=} \bigcap_{i \in \mathcal{C}[s]} \left(\bigcap_{\ell \leq \ell_{\max}(\varepsilon)} C_{i, \ell}(\varepsilon) \cap \mathcal{B}_{i, \ell_{\max}}(\varepsilon) \right)$$

with $\begin{cases} C_{i, \ell}(\varepsilon) \stackrel{\text{def}}{=} \left\{ \left| \mu_i \left(\ell, \frac{\eta}{1 - \eta} U(\ell, \varepsilon) \right) - \mathcal{V}[i] \right| \leq \left(1 + \frac{\eta}{1 - \eta} \right) U(\ell, \varepsilon) \right\} \\ \mathcal{B}_{i, \ell}(\varepsilon) \stackrel{\text{def}}{=} \mathcal{B}_i \left(\ell, \frac{\eta}{1 - \eta} U(\ell, \varepsilon) \right). \end{cases}$

- Else, if s is an **AVG** node,

$$\mathcal{B}_s(m, \varepsilon) \stackrel{\text{def}}{=} \bigcap_{k \in \{1, \dots, m\}} \mathcal{B}_{\tau_{s, k}}(\nu_s(k, m), \varepsilon/\gamma),$$

$$\text{with } \nu_s(k, m) \stackrel{\text{def}}{=} |\{j : \tau_{s, k} = \tau_{s, j} \text{ and } j \leq m\}|.$$

Intuitively, \mathcal{C} is the event where the confidence bounds are satisfied. Using events \mathcal{C} , Definition 3 recursively constructs the events \mathcal{B} needed for Lemma 1.

We remark that there is no loop in this recursive definition as the definition of \mathcal{B}_s only uses $\{\mathcal{B}_i\}_i$ where i is a child of s . Moreover, this way we define \mathcal{B}_s in a finite number of steps as going down through **MAX** and **AVG** nodes increases ε to $\varepsilon\eta/\gamma$ and reaches $1/(1 - \gamma)$ in a finite number of steps.

Moreover, note that \mathcal{B}_s is *not* an intersection of a finite number of $C_{i, \ell}$ because $\tau_{s, k}$ is a random variable and therefore in the **AVG** case, \mathcal{B}_s is not just the intersection of m events.

Remark 4. Notice that, by induction, for any $\delta > 0, \eta > 0, \varepsilon_1, \varepsilon_2, m_1, m_2$, and any node s ,

$$\varepsilon_1 \leq \varepsilon_2 \text{ and } m_1 \geq m_2 \implies \mathcal{B}_s(m_1, \varepsilon_1) \subseteq \mathcal{B}_s(m_2, \varepsilon_2).$$

Lemma 1. For any $k \in \mathbb{N}_0^+$ and any node s , integers $m' \geq m$, and $\varepsilon > 0$, such that $\delta < (\bar{\ell}(m, \eta\varepsilon))^{-h_s(\varepsilon)}$ and

- either k is even, s is an **AVG** node, and $\varepsilon > \frac{(\gamma/\eta)^{k/2}}{1-\gamma}$,
- or k is odd, s is a **MAX** node, and $\varepsilon > \frac{(\gamma/\eta)^{(k-1)/2}}{\eta(1-\gamma)}$,

then the two following properties are verified

Property A: $\mathbb{P}[\mathcal{B}_s(m, \varepsilon)] \geq 1 - \delta \frac{(\bar{\ell}(m, \varepsilon) K/\eta^2)^{h_s(\varepsilon)} - 1}{K - 1}$

Property B: $\forall \lambda, \mathbb{E} \left[e^{\lambda(\mu_s(m, \varepsilon) - \mathcal{V}[s])} | \mathcal{B}_s(m', \varepsilon) \right] \leq \exp \left(2\delta \left(\frac{\bar{\ell}(m, \varepsilon)}{\eta^2} \right)^{h_s(\varepsilon)} \times \frac{1 - \varepsilon\theta(s)}{\eta/\gamma - 1} + \varepsilon|\lambda| + \frac{\lambda^2}{2(1-\gamma)^2 m} \right),$

where $\theta(s) = \begin{cases} 1 - \gamma & \text{if } s \text{ is a MAX node} \\ (1 - \gamma)/\gamma & \text{if } s \text{ is an AVG node} \end{cases}$ and $h_s(\varepsilon) = \max \left(0, \frac{\log(\theta(s)\gamma\varepsilon)}{\log(\gamma/\eta)} \right)$

Lemma 1 claims that on a high-probability event \mathcal{B} , the sub-variance of μ_s is of order at most $1/m$ and its bias of order at most ε up to the term independent of λ . We prove it by induction on k .

Proof. 1° Let $k = 0$. Then, k is even and s is an **AVG** node. Since $\varepsilon > 1/(1 - \gamma)$, s outputs zero by the definition. For any $m' \geq m$ we have $\mathcal{B}_s(m, \varepsilon) = \mathcal{B}_s(m', \varepsilon) = \Omega$, since $\varepsilon > 1/(1 - \gamma)$.

Property A: By definition $h_s(\varepsilon) \geq 0$, therefore $\delta \frac{(\bar{\ell}(m, \varepsilon) K/\eta^2)^{h_s(\varepsilon)} - 1}{K - 1} \geq 0$. Then,

$$\mathbb{P}[\Omega] = 1 \geq 1 - \delta \frac{(\bar{\ell}(m, \varepsilon) K/\eta^2)^{h_s(\varepsilon)} - 1}{K - 1}.$$

Property B: First, as the rewards are bounded from above by one, we have $\mathcal{V}[s] \leq \sum_h \gamma^h \leq 1/(1 - \gamma)$. Second, $\mathcal{V}[s] \geq 0$ as the rewards are bounded from below by zero. Finally, $\varepsilon\theta(s) \leq 1$, hence

$$2\delta \left(\frac{\bar{\ell}(m, \varepsilon)}{\eta^2} \right)^{h_s(\varepsilon)} \times \frac{1 - \varepsilon\theta(s)}{\eta/\gamma - 1} \geq 0.$$

We combine these bounds to get

$$\begin{aligned} \forall \lambda, \mathbb{E} \left[e^{\lambda(\mu_s(m, \varepsilon) - \mathcal{V}[s])} | \mathcal{B}_s(m, \varepsilon) \right] &= \mathbb{E} \left[e^{\lambda(\mu_s(m, \varepsilon) - \mathcal{V}[s])} \right] = \mathbb{E} \left[e^{\lambda(0 - \mathcal{V}[s])} \right] \leq \exp \left(\frac{|\lambda|}{1 - \gamma} \right) \\ &\leq \exp(\varepsilon|\lambda|) \leq \exp \left(2\delta \left(\frac{\bar{\ell}(m, \varepsilon)}{\eta^2} \right)^{h_s(\varepsilon)} \times \frac{1 - \varepsilon\theta(s)}{\eta/\gamma - 1} + \varepsilon|\lambda| + \frac{\lambda^2}{2(1-\gamma)^2 m} \right). \end{aligned}$$

2° Let $k \in \mathbb{N}_0^+$. Assume that Lemma 1 is true for any $k' \leq k - 1$ and let us prove it for k . We prove it separately for k odd and even.

k is odd: Let s be a **MAX** node, m any integer and ε such that

$$\varepsilon > \frac{(\gamma/\eta)^{(k-1)/2}}{\eta(1-\gamma)} \quad \text{and} \quad \frac{1}{\bar{\ell}(m, \varepsilon)} > \delta.$$

We deduce that

$$\eta\varepsilon > \eta \frac{(\gamma/\eta)^{(k-1)/2}}{\eta(1-\gamma)} = \frac{(\gamma/\eta)^{(k-1)/2}}{1-\gamma}.$$

Also, any child of s is an **AVG** node as nodes alternate between **AVG** and **MAX**. We can thus apply the induction assumption to any child of s with parameters (m', ε') with $\varepsilon' < \eta\varepsilon$ and $m' \leq \bar{\ell}(m, \varepsilon)$.

For any child $i \in \mathcal{C}[s]$ of s and any integer $\ell > 0$, we remind the reader the definition of $\mathcal{B}_{i,\ell}$,

$$\mathcal{B}_{i,\ell} = \mathcal{B}_i \left(\ell, \frac{\eta}{1-\eta} \mathbf{U}(\ell, \varepsilon) \right).$$

If $\mathbf{U}(\ell, \varepsilon) \geq (1-\eta)\varepsilon$, then $\frac{\eta}{1-\eta} \mathbf{U}(\ell, \varepsilon) \geq \eta\varepsilon$ and $\ell \leq \ell_{\max}(\varepsilon)$.

We can thus apply our induction assumption to get that for any $\ell \leq \ell_{\max}$,

$$\text{A } \mathbb{P}[\mathcal{B}_{i,\ell}] \geq 1 - \delta \frac{\left(\bar{\ell} \left(\ell, \frac{\eta}{1-\eta} \mathbf{U}(\ell, \varepsilon) \right) K / \eta^2 \right)^{h_i \left(\frac{\eta}{1-\eta} \mathbf{U}(\ell, \varepsilon) \right)} - 1}{K-1} \quad (2)$$

$$\begin{aligned} \text{B } \forall \lambda, \mathbb{E} \left[e^{\lambda(\mu_i(\ell, \frac{\eta}{1-\eta} \mathbf{U}(\ell, \varepsilon)) - \nu[i])} \mathbf{1}_{\mathcal{B}_{i,\ell}} \right] & \quad (3) \\ \leq \exp \left(2\delta \left(\frac{\bar{\ell} \left(\ell, \frac{\eta}{1-\eta} \mathbf{U}(\ell, \varepsilon) \right)}{\eta^2} \right)^{h_i \left(\frac{\eta}{1-\eta} \mathbf{U}(\ell, \varepsilon) \right)} \times \frac{1 - \frac{\eta}{1-\eta} \mathbf{U}(\ell, \varepsilon) \theta(i)}{\eta/\gamma - 1} + \frac{\eta}{1-\eta} \mathbf{U}(\ell, \varepsilon) |\lambda| + \frac{\lambda^2}{2(1-\gamma)^2 \ell} \right) \end{aligned}$$

We first prove Property A.

Combining Inequality 3 with Lemma 4, we get for all $i \in \mathcal{C}[s]$ and $\ell \leq \ell_{\max}$,

$$\begin{aligned} \mathbb{P}[\mathcal{C}_{i,\ell}^c | \mathcal{B}_{i,\ell}] & \leq 2 \exp \left(2\delta \ell_{\max} (\eta\varepsilon)^{h_i(\eta\varepsilon)} \times \frac{1 - \varepsilon \theta(i)}{\eta/\gamma - 1} \right) \exp \left(- \frac{\left(\mathbf{U}(\ell, \varepsilon) \left(1 + \frac{\eta}{1-\eta} \right) - \frac{\eta}{1-\eta} \mathbf{U}(\ell, \varepsilon) \right)^2}{(1-\gamma)^2 \ell} \right) \\ & \leq 2 \exp \left(2 \frac{1 - \varepsilon \theta(i)}{\eta/\gamma - 1} \right) \exp \left(- \frac{(1-\gamma)^2 \ell}{2} \mathbf{U}(\ell, \varepsilon)^2 \right) \\ \text{as } (\delta < 1/\ell) & \leq 2 \exp \left(\frac{2\gamma}{\eta - \gamma} \right) \exp \left(- \frac{(1-\gamma)^2 \ell}{2} \frac{1}{(1-\gamma)^2} \frac{4 \log(K\ell/(\delta\varepsilon(1-\gamma))) + 4\gamma/(\eta - \gamma) + 4}{\ell} \right) \\ & \leq \frac{1}{2} \left(\frac{\delta\varepsilon(1-\gamma)}{K\ell} \right)^2 \leq \frac{1}{2} \frac{\delta\varepsilon(1-\gamma)}{K\ell^2}. \end{aligned}$$

Using a union bound over all $\ell \leq \ell_{\max}$ we get

$$\mathbb{P}[\cup_{\ell \leq \ell_{\max}} \mathcal{C}_{i,\ell}^c | \mathcal{B}_{i,\ell}] \leq \sum_{\ell} \frac{1}{2} \frac{\delta\varepsilon(1-\gamma)}{K\ell^2} \leq \frac{\pi^2}{12} \frac{\delta\varepsilon(1-\gamma)}{K} \leq \frac{\delta\varepsilon(1-\gamma)}{K}.$$

Now, we can finally bound the event for node s as

$$\begin{aligned} \mathbb{P}[\mathcal{B}_s(m, \varepsilon)] & = \mathbb{P}[\cap_{i \in \mathcal{C}[s]} (\cap_{\ell \leq \ell_{\max}} \mathcal{C}_{i,\ell} \cap \mathcal{B}_{i,\ell_{\max}})] \\ & = 1 - \mathbb{P}[\cup_{i \in \mathcal{C}[s]} (\cup_{\ell \leq \ell_{\max}} \mathcal{C}_{i,\ell}^c \cup \mathcal{B}_{i,\ell_{\max}}^c)] \\ & \geq 1 - \sum_{i \in \mathcal{C}[s]} \mathbb{P}[\cup_{\ell \leq \ell_{\max}} \mathcal{C}_{i,\ell}^c \cup \mathcal{B}_{i,\ell_{\max}}^c] \\ & = 1 - \sum_{i \in \mathcal{C}[s]} (1 - \mathbb{P}[\cap_{\ell \leq \ell_{\max}} \mathcal{C}_{i,\ell} \cap \mathcal{B}_{i,\ell_{\max}}]) \\ & = 1 - \sum_{i \in \mathcal{C}[s]} (1 - \mathbb{P}[\cap_{\ell \leq \ell_{\max}} \mathcal{C}_{i,\ell} | \mathcal{B}_{i,\ell_{\max}}] \mathbb{P}[\mathcal{B}_{i,\ell_{\max}}]). \end{aligned}$$

We use Inequality 2 to get

$$\mathbb{P}[\mathcal{B}_{i,\ell_{\max}}] \geq 1 - \delta \frac{\left(\bar{\ell}(\ell_{\max}, \eta\varepsilon) K / \eta^2 \right)^{h_i(\eta\varepsilon)} - 1}{K-1}.$$

To relate $h_s(\varepsilon)$ to its child, we realize that i is an **AVG** node and hence,

$$\begin{aligned}
h_i(\eta\varepsilon) &= \max\left(0, \frac{\log(\theta(s)\gamma\eta\varepsilon(1-\gamma))}{\log(\gamma/\eta)}\right) \\
&= \frac{\log(\theta(s)\gamma\eta\varepsilon(1-\gamma))}{\log(\gamma/\eta)} \\
&= \frac{\log(\eta\varepsilon(1-\gamma))}{\log(\gamma/\eta)} \\
&= \frac{\log(\eta/\gamma) + \log((1/\gamma)\varepsilon(1-\gamma))}{\log(\gamma/\eta)} \\
&= h_s(\varepsilon) - 1.
\end{aligned}$$

We go back to bounding $\mathbb{P}[\mathcal{B}_s(m, \varepsilon)]$,

$$\begin{aligned}
\mathbb{P}[\mathcal{B}_s(m, \varepsilon)] &\geq 1 - \sum_{i \in \mathcal{C}[s]} \left(1 - \left(\frac{\delta\varepsilon(1-\gamma)}{K}\right) \left(1 - \delta \frac{(\bar{\ell}(\ell_{\max}, \eta\varepsilon) K/\eta^2)^{h_i(\eta\varepsilon)} - 1}{K-1}\right)\right) \\
&\geq 1 - \sum_{i \in \mathcal{C}[s]} \left(\frac{\delta}{K} + \delta \frac{(\bar{\ell}(\ell_{\max}, \varepsilon) K)^{h_s(\varepsilon)-1} - 1}{K-1}\right) \quad (\text{as } \varepsilon \leq 1/(1-\gamma)) \\
&= 1 - \delta \left(1 + K \frac{(\bar{\ell}(\ell_{\max}, \varepsilon) K)^{h_s(\varepsilon)-1} - 1}{K-1}\right) \\
&\geq 1 - \delta \frac{(\bar{\ell}(m, \varepsilon) K/\eta^2)^{h_s(\varepsilon)} - 1}{K-1}.
\end{aligned}$$

To prove Property **B**, we introduce additional notation. We define

$$\mu_{i,\ell} \stackrel{\text{def}}{=} \mu_i\left(\ell, \frac{\eta}{1-\eta} \mathcal{U}(\ell, \varepsilon)\right).$$

We also define the empirically best and the best child as

$$\hat{i}_\ell \stackrel{\text{def}}{=} \arg \max_{i \in \mathcal{C}[s]} \mu_{i,\ell} \quad \text{and} \quad i^* \stackrel{\text{def}}{=} \arg \max_{i \in \mathcal{C}[s]} \mathcal{V}[i].$$

We define the random variable \mathcal{L}_ℓ as the set of all children that would be kept in Line 9 in Figure 3 if they are called with parameter ℓ as

$$\mathcal{L}_\ell \stackrel{\text{def}}{=} \left\{i \in \mathcal{C}[s] : \mu_{i,\ell} + \frac{2\mathcal{U}(\ell, \varepsilon)}{1-\eta} \geq \sup_j \left[\mu_{j,\ell} - \frac{2\mathcal{U}(\ell, \varepsilon)}{1-\eta}\right]\right\}.$$

Notice that the **MAX** node is not constructing sets $\{\mathcal{L}_\ell\}_\ell$ iteratively: The reason is that some of the children might have been eliminated in the earlier rounds. However, at the end of the while loop in Figure 3, we get set \mathcal{L} that is the intersection of all $\{\mathcal{L}_\ell\}_\ell$,

$$\mathcal{L} = \bigcap_{\ell \leq \ell_{\max}} \mathcal{L}_\ell.$$

We define \mathcal{G}_ℓ (which is *not* a random variable) as

$$\mathcal{G}_\ell \stackrel{\text{def}}{=} \left\{i \in \mathcal{C}[s] : \mathcal{V}[i] + \frac{\mathcal{U}(\ell, \varepsilon)}{1-\eta} \geq \mathcal{V}[i^*] - \frac{\mathcal{U}(\ell, \varepsilon)}{1-\eta}\right\}.$$

For simplicity, we also define $\hat{i} \stackrel{\text{def}}{=} \hat{i}_{\ell_{\max}}$, $\mathcal{G} \stackrel{\text{def}}{=} \mathcal{G}_{\ell_{\max}}$, and for all $i \in \mathcal{C}[s]$, $\mu_i \stackrel{\text{def}}{=} \mu_{i,\ell_{\max}}$.

We can now prove Property **B**. By definition of $\mathcal{C}_{i,\ell}$ and induction assumption, we have

$$\{\bigcap_{i \in \mathcal{C}[s], \ell \leq \ell_{\max}} \mathcal{C}_{i,\ell}\} \subseteq \{\forall i, \ell : |\mu_{i,\ell} - \mathcal{V}[i]| \leq \mathcal{U}(\ell, \varepsilon)/(1-\eta)\}$$

and we set $\mathcal{C} \stackrel{\text{def}}{=} \bigcap_{i \in \mathcal{C}[s], \ell \leq \ell_{\max}} \mathcal{C}_{i,\ell}$.

We distinguish two cases depending on the cardinality of the *deterministic* set \mathcal{G} and assume \mathcal{C} .

First case: $|\mathcal{G}| = 1$. If $|\mathcal{L}| > 1$ then for all $i \in \mathcal{L}$ such that $i \notin \mathcal{G} = \{i^*\}$

$$\begin{aligned} \mu_i &\leq \mathcal{V}[i] + \varepsilon && \text{(because } C_{i, \ell_{\max}} \text{ holds)} \\ &< V(i^*) - \varepsilon && \text{(because } i \notin \mathcal{G}_s = \{i^*\} \text{)} \\ &\leq \mu_{i^*}^{s, t} && \text{(because } C_{i^*, \ell_{\max}} \text{ holds)} \end{aligned}$$

Therefore $\arg \max_{i \in \mathcal{L}} \mu_i = i^*$ which implies that i^* is called with parameters (m', ε') with

$$\varepsilon' \leq \eta\varepsilon \leq \varepsilon \quad \text{and} \quad m' \geq m.$$

If $|\mathcal{L}| = 1$, then $\mathcal{L} = \{i^*\}$ and consequently i^* is called with parameters (m, ε) .

In both cases, on event C , $\mu_s(m, \varepsilon)$ is the output of a call to i^* with parameters (m', ε') with $m' \geq m$ and $\varepsilon' \leq \varepsilon$ and $m' \leq \ell_{\max}$. We use Property **B** of induction assumption as $\ell_{\max} \geq m$ and $U(\ell_{\max}, \varepsilon)\eta/(1-\eta) \leq \varepsilon$.

$$\begin{aligned} \forall \lambda, \mathbb{E} \left[e^{\lambda(\mu_{i^*}(m', \varepsilon') - \mathcal{V}[i^*])} \middle| B_{i^*, \ell_{\max}} \right] &= \mathbb{E} \left[e^{\lambda(\mu_s(m, \varepsilon) - \mathcal{V}[s])} \middle| B_{i^*, \ell_{\max}} \right] \\ &\leq \exp \left(2\delta \left(\frac{\bar{\ell}(m', \varepsilon')}{\eta^2} \right)^{h_{i^*}(\varepsilon')} \times \frac{1 - \varepsilon'\theta(i^*)}{\eta/\gamma - 1} + \varepsilon'|\lambda| + \frac{\lambda^2}{2(1-\gamma)^2 m'} \right) \\ &\leq \exp \left(2\delta \left(\frac{\bar{\ell}(m, \eta\varepsilon)}{\eta^2} \right)^{h_s(\eta\varepsilon)} \times \frac{1 - \eta\varepsilon\theta(s)}{\eta/\gamma - 1} + \varepsilon|\lambda| + \frac{\lambda^2}{2(1-\gamma)^2 m} \right) \end{aligned}$$

Because for all $i, j \in \mathcal{C}[s]$ with $i \neq j$, B_i and μ_j are independent, we can claim that

$$\forall \lambda, \mathbb{E} \left[e^{\lambda(\mu_s(m, \varepsilon) - \mathcal{V}[s])} \middle| \bigcap_{i \in \mathcal{C}[s]} B_{i, \ell_{\max}} \right] \leq \exp \left(2\delta \left(\frac{\bar{\ell}(m, \eta\varepsilon)}{\eta^2} \right)^{h_s(\eta\varepsilon)} \times \frac{1 - \eta\varepsilon\theta(s)}{\eta/\gamma - 1} + \varepsilon|\lambda| + \frac{\lambda^2}{2(1-\gamma)^2 m} \right).$$

Using Lemma 5 we finally get for all λ

$$\begin{aligned} \mathbb{E} \left[e^{\lambda(\mu_s(m, \varepsilon) - \mathcal{V}[s])} \middle| \mathcal{B}_s(m, \varepsilon) \right] &= \mathbb{E} \left[e^{\lambda(\mu_s(m, \varepsilon) - \mathcal{V}[s])} \middle| \bigcap_{i \in \mathcal{C}[s]} (\bigcap_{\ell \leq \ell_{\max}} C_{i, \ell} \cap B_{i, \ell_{\max}}) \right] \\ &\leq \frac{\mathbb{E} \left[e^{\lambda(\mu_s(m, \varepsilon) - \mathcal{V}[s])} \middle| \bigcap_{i \in \mathcal{C}[s]} B_{i, \ell_{\max}} \right]}{\mathbb{P} \left[\bigcap_{i \in \mathcal{C}[s], \ell \leq \ell_{\max}} C_{i, \ell} \middle| \bigcap_{i \in \mathcal{C}[s]} B_{i, \ell_{\max}} \right]} \\ &\leq \frac{\exp \left(2\delta \left(\frac{\bar{\ell}(m, \eta\varepsilon)}{\eta^2} \right)^{h_s(\eta\varepsilon)} \times \frac{1 - \eta\varepsilon\theta(s)}{\eta/\gamma - 1} + \varepsilon|\lambda| + \frac{\lambda^2}{2(1-\gamma)^2 m} \right)}{1 - \delta\varepsilon(1-\gamma)} \end{aligned}$$

To simplify the expression, we note that for $0 < x \leq 1/2$, we have that $1/(1-x) \leq 1+2x \leq \exp(2x)$. We also use that $\delta < 1/2$.

$$\begin{aligned} \frac{\exp \left(2\delta \left(\frac{\bar{\ell}(m, \eta\varepsilon)}{\eta^2} \right)^{h_s(\eta\varepsilon)} \times \frac{1 - \eta\varepsilon\theta(s)}{\eta/\gamma - 1} \right)}{1 - \delta\varepsilon(1-\gamma)} &\leq \exp \left(2\delta \left(\frac{\bar{\ell}(m, \eta\varepsilon)}{\eta^2} \right)^{h_s(\eta\varepsilon)} \times \frac{1 - \eta\varepsilon\theta(s)}{\eta/\gamma - 1} + 2\delta\varepsilon(1-\gamma) \right) \\ &\leq \exp \left(2\delta \left(\frac{\bar{\ell}(m, \varepsilon)}{\eta^2} \right)^{h_s(\varepsilon)} \left(\frac{1 - \eta\varepsilon(1-\gamma)/\gamma}{\eta/\gamma - 1} + \varepsilon(1-\gamma) \right) \right) \\ &= \exp \left(2\delta \left(\frac{\bar{\ell}(m, \varepsilon)}{\eta^2} \right)^{h_s(\varepsilon)} \frac{1 - \varepsilon(1-\gamma)}{\eta/\gamma - 1} \right) \\ &= \exp \left(2\delta \left(\frac{\bar{\ell}(m, \varepsilon)}{\eta^2} \right)^{h_s(\varepsilon)} \frac{1 - \varepsilon\theta(s)}{\eta/\gamma - 1} \right) \end{aligned}$$

Second case: $|\mathcal{G}| > 1$. For all $i \in \mathcal{G}_l$,

$$\begin{aligned} \mu_{i,\ell} &\geq \mathcal{V}[i] - \mathsf{U}(\ell, \varepsilon)/(1 - \eta) && \text{(because } C \text{ holds)} \\ &\geq V(i^*) - 3\mathsf{U}(\ell, \varepsilon)/(1 - \eta) && \text{(because } i \in \mathcal{G}_l) \\ &\geq \mu_{i^*}^{s,t} - 4\mathsf{U}(\ell, \varepsilon)/(1 - \eta) && \text{(because } C \text{ holds)} \end{aligned}$$

Therefore for all $\ell \leq \ell_{\max}$,

$$\mathcal{G}_\ell \subset \mathcal{L}_\ell.$$

As a result $|\mathcal{L}| > |\mathcal{G}| > 1$. The output is then the maximum of the estimates in \mathcal{L} . The best estimate $\mu_{\hat{i}}$ in \mathcal{L} verifies

$$\begin{aligned} \mu_{\hat{i}} &\geq \mu_{i^*} && \text{(by definition of } \hat{i}) \\ &\geq V(i^*) - \mathsf{U}(\ell_{\max}, \varepsilon)/(1 - \eta) && \text{(because } C \text{ holds)}. \end{aligned}$$

It also verifies that

$$\begin{aligned} \mu_{\hat{i}} &\leq V(\hat{i}) + \mathsf{U}(\ell_{\max}, \varepsilon)/(1 - \eta) && \text{(because } C_{\hat{i}, \ell_{\max}} \text{ holds)} \\ &\leq V(i^*) + \mathsf{U}(\ell_{\max}, \varepsilon)/(1 - \eta) && \text{(by definition of } i^*). \end{aligned}$$

Since $\mu_s(m, \varepsilon) = \mu_{\hat{i}}$, we have

$$|\mu_s(m, \varepsilon) - \mathcal{V}[s]| \leq \frac{\mathsf{U}(\ell_{\max}, \varepsilon)}{1 - \eta} \leq \varepsilon.$$

It follows that

$$\begin{aligned} \mathbb{E} \left[e^{\lambda(\mu_s(m, \varepsilon) - \mathcal{V}[s])} | \mathcal{B}_s(m, \varepsilon) \right] &\leq \exp(|\lambda|\varepsilon) \\ &\leq \exp \left(2\delta \left(\frac{\bar{\ell}(m, \varepsilon, \cdot)}{\eta^2} \right)^{h_s(\varepsilon, \cdot)} \times \frac{1 - \varepsilon, \theta(s)}{\eta/\gamma - 1} + |\lambda|\varepsilon + \frac{\lambda^2}{2(1 - \gamma)^2 m} \right). \end{aligned}$$

k is even: Let s be a **AVG** node, $m \in \mathbb{N}_0^+$, and ε such that

$$\varepsilon > \frac{(\gamma/\eta)^{k/2}}{(1 - \gamma)} \quad \text{and} \quad \frac{1}{\max(m, \ell_{\max}(\varepsilon))} > \delta.$$

Hence, we can deduce

$$\frac{\varepsilon}{\gamma} > \frac{(\gamma/\eta)^{k/2}}{\gamma(1 - \gamma)} = \frac{(\gamma/\eta)^{(k-2)/2}}{\eta(1 - \gamma)}.$$

Also, any child of s is a **MAX** node as nodes alternate between **AVG** and **MAX**. We can thus apply induction assumption to any child of s with parameters (m', ε') with $\varepsilon' \geq \varepsilon/\gamma$ and $m' \leq \max(m, \ell_{\max}(\varepsilon))$.

For any child $i \in \mathcal{C}[s]$ of s and any integer $\ell \leq m$ we can thus apply our induction assumption to get

$$\begin{aligned} \mathbb{P}[\mathcal{B}_i(\ell, \varepsilon/\gamma)] &\geq 1 - \delta \frac{(\bar{\ell}(\ell, \varepsilon/\gamma) K/\eta^2)^{h_i(\varepsilon/\gamma)} - 1}{K - 1} \\ \forall \lambda, \mathbb{E} \left[e^{\lambda(\mu_i(\ell, \varepsilon/\gamma) - \mathcal{V}[i])} | \mathcal{B}_i(\ell, \varepsilon/\gamma) \right] &\leq \exp \left(2\delta \left(\frac{\bar{\ell}(\ell, \varepsilon/\gamma)}{\eta^2} \right)^{h_i(\varepsilon/\gamma)} \times \frac{1 - \varepsilon/\gamma \theta(i)}{\eta/\gamma - 1} + |\lambda|\varepsilon/\gamma + \frac{\lambda^2}{2(1 - \gamma)^2 \ell} \right) \end{aligned}$$

We state again the definition of \mathcal{B}_s

$$\mathcal{B}_s(m, \varepsilon) \stackrel{\text{def}}{=} \bigcap_{i \in \mathcal{C}[s]} \mathcal{B}_i(\nu_i(m), \varepsilon/\gamma)$$

with $\nu_i(m) \stackrel{\text{def}}{=} |\{j : \tau_{s,j} = i \text{ and } j \leq m\}|$

We first prove property **A**.

$$\begin{aligned}
\mathbb{P} \left[\bigcup_{i \in \mathcal{C}[s]} \mathcal{B}_i(\nu_i(m), \varepsilon/\gamma)^c \right] &\leq \sum_{i \in \mathcal{C}[s]} \mathbb{P} [\mathcal{B}_i(\nu_i(m), \varepsilon/\gamma)^c] \\
&= \int_k \sum_{i \in \mathcal{C}[s]} \mathbb{P} [\mathcal{B}_i(\nu_i(m), \varepsilon/\gamma)^c | \nu_i(m) = k] d\pi(k) \\
&= \int_k \sum_{i \in \mathcal{C}[s]} \mathbb{P} [\mathcal{B}_i(k_i, \varepsilon/\gamma)^c | \nu_i(m) = k_i] d\pi(k) \\
&= \int_k \sum_{i \in \mathcal{C}[s]} \mathbb{P} [\mathcal{B}_i(k_i, \varepsilon/\gamma)^c] d\pi(k) \\
&\leq \int_k \sum_{i \in \mathcal{C}[s]; k_i > 0} \delta \frac{(\bar{\ell}(k_i, \varepsilon/\gamma) K/\eta^2)^{h_i(\varepsilon/\gamma)} - 1}{K-1} \pi(k) \\
&\leq \int_k \delta \frac{(\bar{\ell}(k_i, \varepsilon) K/\eta^2)^{h(\varepsilon)} - 1}{K-1} \pi(k) \\
&= \delta \frac{(\bar{\ell}(k_i, \varepsilon) K/\eta^2)^{h(\varepsilon)} - 1}{K-1}
\end{aligned}$$

We now prove property **B**.

$$\begin{aligned}
&\mathbb{E} \left[\exp \left((\lambda/m) \sum_{i \in \mathcal{C}[s]} \nu_i(m) \mu_i(\nu_i(m), \varepsilon/\gamma) \right) \middle| \mathcal{B}_s(m, \varepsilon) \right] \\
&= \int_k \mathbb{E} \left[\exp \left((\lambda/m) \sum_{i \in \mathcal{C}[s]} \nu_i(m) \mu_i(\nu_i(m), \varepsilon/\gamma) \right) \middle| \mathcal{B}_s(m, \varepsilon), (\nu_i(m)) = (k_i) \right] d\pi(k) \\
&= \int_k \mathbb{E} \left[\exp \left((\lambda/m) \sum_{i \in \mathcal{C}[s]} k_i \mu_i(k_i, \varepsilon/\gamma) \right) \middle| \mathcal{B}_s(m, \varepsilon) \right] d\pi(k) \\
&= \int_k \mathbb{E} \left[\prod_{i \in \mathcal{C}[s]} \exp((\lambda/m) k_i \mu_i(k_i, \varepsilon/\gamma)) \middle| \mathcal{B}_s(m, \varepsilon) \right] d\pi(k) \\
&\leq \int_k \exp \left(\sum_{i \in \mathcal{C}[s]} k_i \left(2\delta \left(\frac{\bar{\ell}(k_i, \varepsilon/\gamma)}{\eta^2} \right)^{h_i(\varepsilon/\gamma)} \times \frac{1 - \varepsilon/\gamma \theta(i)}{\eta/\gamma - 1} + |(\lambda/m)\varepsilon/\gamma + \frac{(\lambda/m)^2}{2(1-\gamma)^2 k_i}| \right) \right) d\pi(k) \\
&= \int_k \exp \left(m 2\delta \left(\frac{\bar{\ell}(m, \varepsilon/\gamma)}{\eta^2} \right)^{h_s(\varepsilon/\gamma)} \times \frac{1 - \varepsilon/\gamma \theta(i)}{\eta/\gamma - 1} + |\lambda|\varepsilon/\gamma + \frac{\lambda^2}{2(1-\gamma)^2 m} \right) d\pi(k) \\
&\leq \exp \left(2\delta \left(\frac{\bar{\ell}(m, \varepsilon)}{\eta^2} \right)^{h_s(\varepsilon)} \times \frac{1 - \varepsilon \theta(s)}{\eta/\gamma - 1} + |\lambda|\varepsilon/\gamma + \frac{\lambda^2}{2(1-\gamma)^2 m} \right)
\end{aligned}$$

We can finally bound

$$\begin{aligned}
& \mathbb{E} \left[\exp \left(\lambda r + \gamma(\lambda/m) \sum_{i \in \mathcal{C}[s]} \nu_i(m) \mu_i(\nu_i(m), \varepsilon/\gamma) \right) \middle| \mathcal{B}_s(m, \varepsilon) \right] \\
& \leq \exp \left(\frac{\lambda^2}{2m} + 2\delta \left(\frac{\bar{\ell}(m, \varepsilon)}{\eta^2} \right)^{h_s(\varepsilon)} \times \frac{1 - \varepsilon\theta(s)}{\eta/\gamma - 1} + |\lambda|\varepsilon + \frac{\lambda^2 \gamma^2}{2(1 - \gamma)^2 m} \right) \\
& \leq \exp \left(2\delta \left(\frac{\bar{\ell}(m, \varepsilon)}{\eta^2} \right)^{h_s(\varepsilon)} \times \frac{1 - \varepsilon\theta(s)}{\eta/\gamma - 1} + |\lambda|\varepsilon + \left(\frac{\lambda\gamma}{\sqrt{2m(1 - \gamma)}} + \frac{\lambda}{\sqrt{2m}} \right)^2 \right) \\
& = \exp \left(2\delta \left(\frac{\bar{\ell}(m, \varepsilon)}{\eta^2} \right)^{h_s(\varepsilon)} \times \frac{1 - \varepsilon\theta(s)}{\eta/\gamma - 1} + |\lambda|\varepsilon + \frac{\lambda^2}{2(1 - \gamma)^2 m} \right)
\end{aligned}$$

□

Now we stat and prove Theorem 4 that is equivalent to Theorem 1.

Theorem 4. *Let μ be the output of TrailBlazer. For all $\delta' > 0, \varepsilon > 0$, there exists an event \mathcal{B} such that $\mathbb{P}[\mathcal{B}] \geq 1 - \delta'$ and on \mathcal{B} , $|\mu - \mathcal{V}[s_0]| \leq \varepsilon$.*

Proof. We apply Lemma 1 to the root node s_0 . Let's note

$$\begin{aligned}
m &= (\log(1/\delta) + \lambda)/((1 - \gamma)^2 \varepsilon^2) \\
\lambda &= 2 \log(\varepsilon(1 - \gamma))^2 \frac{\log\left(\frac{\log(K)}{(1 - \eta)}\right)}{\log(\gamma/\eta)}
\end{aligned}$$

We set $\delta = \exp(\lambda)\delta'$. We apply Lemma 1 on s_0 with $(m, \varepsilon/2)$. We take $\mathcal{B} = \mathcal{B}_{s_0}(\log(1/\delta) + \lambda)/((1 - \gamma)^2 \varepsilon^2)$.

We check that $\delta \frac{(\bar{\ell}(m, \varepsilon)K/\eta^2)^{h_s(\varepsilon)} - 1}{K - 1} \leq \delta'$ and apply Lemma 4 with property B of Lemma 1 to get $|\mu_{s_0}(m, \varepsilon/2) - \mathcal{V}[s_0]| \leq \varepsilon$. □

B Sample complexity

Definition 4 (node classification). A node s of depth h is said to be

- sub-optimal if there exists $h' < h$ such that $\Delta_{\rightarrow s}(s_{h'}) > 16 \frac{\gamma^{h-h'}}{\gamma(1-\gamma)}$,
- optimal if for all $h' < h$ $\Delta_{\rightarrow s}(s_{h'}) = 0$,
- near-optimal in any other cases.

Lemma 2. For any node s of depth h , integer m , and $\varepsilon > 0$, a call to s with parameters (m, ε) does not generate call to nodes of depth $h' > h + h_{\max}(\varepsilon)$ with

$$h_{\max}(\varepsilon) = \left\lceil \frac{\log(\varepsilon(1-\gamma))}{\log(\gamma/\eta)} \right\rceil.$$

Lemma 3. If $\mathcal{B}_{s_0}(m, \varepsilon)$ holds then for all nodes s of depth h and $h' < h$, the following holds,

$$s \text{ is suboptimal} \implies \Gamma_s(m, \varepsilon) = 0$$

$$\Delta(s_{h'}) > 16 \frac{\gamma^{h-h'}}{\gamma(1-\gamma)} \implies \text{any call to } s \text{ was generated by a call from } s_{h'} \text{ with parameter } m = \Gamma_{s_{h'}}(m_0, \varepsilon_0).$$

Proof. We denote the parameters the root has been called with by (m_0, ε_0) . For any node s of depth h , we assume that $\Gamma_s(m_0, \varepsilon_0) > 0$ and we want to prove that s is not sub-optimal.

Let h' be some integer with $h' < h$. We note $(m_{s_{h'}}, \varepsilon_{s_{h'}})$ the parameters $s_{h'}$ has been called with. If $\Gamma_s(m_0, \varepsilon_0) > 0$, then, by Lemma 2, that means that $s_{h'}$ performs a call to its child i_s in direction to s with parameters (m, ε) such that $h_{\max}(\varepsilon) \geq h - h'$.

For any m', ε' , if $s_{h'}$ performs a call to i_s with parameters (m', ε') with $\varepsilon' \leq \varepsilon$ then

1. either $i_s \in \mathcal{L}_{\ell_{\max}(m_{s_{h'}}, \varepsilon_{s_{h'}})}$ and $\eta \varepsilon_{s_{h'}} \leq \varepsilon$,
2. or $i_s \in \mathcal{L}_\ell$ for some ℓ such that $\frac{\eta}{1-\eta} \mathbb{U}_\ell \leq \varepsilon$.

Case 1. For all ℓ' we have $\mathcal{L}_{\ell'} \subset \mathcal{H}_{\ell'}$ therefore $V(i^*) - V(i_s) \leq \frac{4\mathbb{U}_{\ell'}}{1-\eta} \leq 4\frac{\varepsilon}{\eta}$.

Case 2. By definition of $\ell_{\max} : \mathbb{U}_{\ell_{\max}+1} \leq (1-\eta)\varepsilon_{s_{h'}}$. We also have

$$\forall \ell' \exists a \geq 0 \quad \frac{\mathbb{U}_{\ell'+1}}{\mathbb{U}_{\ell'}} = \sqrt{\frac{\ell' \log(\ell'+1) + a}{\ell' \log(\ell') + a}} \geq \sqrt{\frac{\ell'}{\ell'+1}} \geq 1/\sqrt{2}.$$

We put it together using $i_s \in \mathcal{L}_{\ell_{\max}} \subset \mathcal{H}_{\ell_{\max}}$,

$$V(i^*) - V(i_s) \leq \frac{4\mathbb{U}_{\ell_{\max}}}{1-\eta} \leq \mathbb{U}_{\ell_{\max}} \leq \frac{4\sqrt{2}}{1-\eta} \mathbb{U}_{\ell_{\max}+1} \leq 4\sqrt{2}\varepsilon_s \leq 4\sqrt{2}\varepsilon/\eta.$$

Either $V(i^*) - V(i_s) \leq \frac{4\varepsilon}{\eta}$ or $V(i^*) - V(i_s) \leq \frac{4\varepsilon\sqrt{2}}{\eta}$ from which we deduce that

$$V(i^*) - V(i_s) \leq \frac{4\varepsilon\sqrt{2}}{\eta}.$$

We use $h - h' \leq h_{\max}(\varepsilon) \leq \frac{\varepsilon(1-\gamma)}{\log(\gamma/\eta)} + 1$ to get $(\gamma/\eta)^{h-h'} \geq \varepsilon(1-\gamma)\gamma/\eta$ and using $\eta \leq 1$

$$\varepsilon \leq \frac{1}{\gamma(1-\gamma)} (1/\eta)^{h-h'} \gamma^{h-h'}.$$

As $h - h' \leq h_{\max}(\varepsilon_0)$,

$$\log\left((1/\eta)^{h-h'}\right) \leq \log\left((1/\eta)^{h_{\max}(\varepsilon_0)}\right) \leq \frac{\log 1/\gamma}{\log(1/\varepsilon)} \frac{\log(1/\varepsilon)}{\log(\eta/\gamma)} \leq 1.$$

We combine

$$\varepsilon \leq \frac{e}{\gamma(1-\gamma)} \gamma^{h-h'} \text{ and } V(i^*) - V(i_s) \leq \frac{4\varepsilon\sqrt{2}}{\eta},$$

to finally get

$$\Delta_{\rightarrow s}(s_{h'}) = V(i^*) - V(i_s) \leq \frac{4e\sqrt{2}}{\gamma(1-\gamma)} \gamma^{h-h'} \leq \frac{16}{\gamma(1-\gamma)} \gamma^{h-h'}.$$

This is true for any $h' < h$ which proves that s is not suboptimal.

We now assume that $\Delta(s_{h'}) > \frac{16}{\gamma(1-\gamma)} \gamma^{h-h'}$ and prove that any call to i^* generating calls to s are done with parameters (m, ε) with $m = \Gamma_{s_{h'}}(m_0, \varepsilon_0)$.

By Lemma 2, $h_{\max}(\varepsilon) \geq h - h'$. We distinguish three cases,

1. either $s_{h'}$ calls i^* within the while loop,
2. or $s_{h'}$ calls i^* outside the while loop with $|\mathcal{L}| > 1$,
3. or $s_{h'}$ calls i^* outside the while loop with $|\mathcal{L}| = 1$.

Case 1. If this call is made to $i \neq i^*$ with $\varepsilon = \frac{\eta}{1-\eta} U_\ell$, with the same reasoning as above we conclude that $V(i^*) - V(i_s) \leq \frac{4e\sqrt{2}}{\gamma(1-\gamma)} \gamma^{h-h'}$ which contradicts the assumption.

Case 2. We use again the same reasoning as above with $U_{\ell_{\max}}$ to conclude that if $\varepsilon_{s_{h'}} \geq \eta\varepsilon$ then $V(i^*) - V(i_s) \leq \frac{4e\sqrt{2}}{\gamma(1-\gamma)} \gamma^{h-h'}$.

Case 3. In this last case, the call is done with parameters $(m_{s_{h'}}, \varepsilon_{s_{h'}})$ and $m_{s_{h'}} \leq \Gamma_{s_{h'}}(m_0, \varepsilon_0)$. □

B.1 Proof of Theorem 3

Theorem 5. *The number of calls T to the model verifies*

$$\mathbb{E}[T] \leq \mathcal{O}\left(\frac{(\log(1/\delta) + \log(1/\varepsilon))^3}{\varepsilon^2}\right).$$

Proof. For any MAX node s' of depth h' and any $h > h'$ we define $\Gamma_{s'}^h$ as the maximum m the node s' has been called with such that this call generated call up to depth h . We first prove that for any γ there exists ξ_1, ξ_2 such that for any node s' of depth $h' < h$ any $i \in \cup_{j \in \mathcal{C}[s']} \mathcal{C}[j]$ and any ε, δ ,

$$\mathbb{E}[\Gamma_i^h] \leq \mathbb{E}[\Gamma_{s'}^h] p(i|s) x^{1(\Delta_s < y)},$$

with $x_{h'} \leq \xi_1 (1/\gamma)^{2(h-h')}$ and $y_{h'} \leq \xi_2 \gamma^{h-h'}$.

If $\Delta_{s'} \geq 16 \frac{\gamma^{h-h'}}{\gamma(1-\gamma)}$ then by Lemma 3, the calls towards i of s' which generate calls up to depth h are done with parameter $m < \mathbb{E}[\Gamma_{s'}^h]$. Thus,

$$\mathbb{E}[\Gamma_i^h] \leq \mathbb{E}[\Gamma_{s'}^h] p(i|s).$$

In the other case, a node called with parameter (m, ε) generates at most $\mathcal{O}(1/\varepsilon^2)$ samples. Moreover, since s' is of depth h' , we know that $\varepsilon \leq \mathcal{O}(\gamma^{h-h'})$ and therefore,

$$\begin{aligned} \mathbb{E}[\Gamma_i^h] &\leq \mathbb{P}[\Gamma_{s'}^h \geq 1] p(i|s) \xi_1 \gamma^{-2(h-h')} \\ &\leq \mathbb{E}[\Gamma_{s'}^h] p(i|s) \xi_1 \gamma^{-2(h-h')}. \end{aligned}$$

From $\mathbb{E} [\Gamma_i^h] \leq \mathbb{E} [\Gamma_s^h] p(i|s)x^{\mathbb{1}(\Delta_s < y)}$, we deduce that for all node s of depth h ,

$$\begin{aligned}
\mathbb{E}_{\Delta_s} [\mathbb{E} [\Gamma_s]] &= \mathbb{E}_{\Delta_s} [\mathbb{E} [\Gamma_s^h]] \leq \mathbb{E}_{\Delta_s} \left[\mathbb{E} [\Gamma_{s_0}^h] p(s|s_0) \prod_{h'=1}^h x_{h'}^{\mathbb{1}(\Delta_{s_{h'}} < y_{h'})} \right] \\
&= \mathbb{E} [\Gamma_{s_0}^h] p(s|s_0) \prod_{h'=1}^h \mathbb{E}_{\Delta_s} [x_{h'}^{\mathbb{1}(\Delta_{s_{h'}} < y_{h'})}] \\
&\leq \mathbb{E} [\Gamma_{s_0}^h] p(s|s_0) \prod_{h'=1}^h [1 + x_{h'} \mathbb{P} [\Delta_{s_{h'}} < y_{h'}]] \\
&\leq \mathbb{E} [\Gamma_{s_0}^h] p(s|s_0) \prod_{h'=1}^h [1 + ax_{h'} y_{h'}^d] \\
&\leq \mathbb{E} [\Gamma_{s_0}^h] p(s|s_0) \exp \left(\sum_{h'=1}^h ax_{h'} y_{h'}^d \right) \\
&\leq \mathbb{E} [\Gamma_{s_0}^h] p(s|s_0) \exp \left(\sum_{h'=1}^h a\xi_1 (1/\gamma)^{2(h-h')} (\xi_2 \gamma^{h'-h})^d \right) \\
&\leq \xi_3 \mathbb{E} [\Gamma_{s_0}^h] p(s|s_0) \exp \left(\sum_{h'=1}^{\infty} \gamma^{(d-2)(h-h')} \right) \\
&\leq \xi_4 \mathbb{E} [\Gamma_{s_0}^h] p(s|s_0).
\end{aligned}$$

Finally,

$$\begin{aligned}
\mathbb{E}_{\Delta_s} \left[\sum_{h=1}^{h_{\max}(\varepsilon)} \sum_{s \text{ of depth } h} \mathbb{E} [\Gamma_s] \right] &\leq \sum_{h=1}^{h_{\max}(\varepsilon)} \sum_{s \text{ of depth } h} \xi_4 \mathbb{E} [\Gamma_{s_0}^h] p(s|s_0) \\
&= \sum_{h=1}^{h_{\max}(\varepsilon)} \xi_4 \mathbb{E} [\Gamma_{s_0}^h] \\
&\leq \xi_4 h_{\max}(\varepsilon) \mathbb{E} [\Gamma_{s_0}^{h_{\max}(\varepsilon)}] \\
&\leq \xi_5 \log(1/\varepsilon) \frac{\log(1/\delta) + \log(1/\varepsilon)}{\varepsilon^2} \\
&\leq \mathcal{O} \left(\frac{(\log(1/\delta) + \log(1/\varepsilon))^3}{\varepsilon^2} \right).
\end{aligned}$$

This is true on $\mathcal{B}_s(m, \varepsilon)$ which holds with probability $1 - \delta$ but as $\delta < \mathcal{O}(1/\varepsilon^3)$, the result still holds on the whole probability space. \square

B.2 Proof of Theorem 2

Proof. Any near-optimal node of depth h called with parameters (m', ε') generates at most $\frac{\log(1/(\varepsilon\delta))}{(1-\eta)^2(1-\gamma)^2\varepsilon'^2}$ samples. This follows from the same reasoning as in previous theorem. Then we can use that $\varepsilon' \leq \varepsilon(\eta/\gamma)^h$.

Therefore, the maximum number of samples T is

$$T \leq \sum_{h \leq h_{\max}} N^h \kappa^h \frac{\log(1/(\varepsilon\delta))}{(1-\eta)^2(1-\gamma)^2(\varepsilon\eta^h/\gamma^h)^2}.$$

If $N\kappa/\gamma^2 > 1$, then,

$$\begin{aligned}
T &\leq \mathcal{O} \left(h_{\max}(N\kappa\eta^2/\gamma^2)^{h_{\max}} \frac{\log(1/(\varepsilon\delta))}{(1-\eta)^2(1-\gamma)^2\varepsilon^2} \right) \\
&= \mathcal{O} \left(h_{\max}(1/\varepsilon)^{(\log(N\kappa)-2\log(\eta/\gamma))/\log(\eta/\gamma)} \frac{\log(1/(\varepsilon\delta))}{(1-\eta)^2(1-\gamma)^2\varepsilon^2} \right) \\
&= \mathcal{O} \left(\frac{\log(1/\varepsilon)}{\log(\eta/\gamma)} (1/\varepsilon)^{\log(N\kappa)/\log(\eta/\gamma)} \frac{\log(1/(\varepsilon\delta))}{(1-\eta)^2(1-\gamma)^2} \right) \\
&= \mathcal{O} \left((1/\varepsilon)^{\log(N\kappa)/\log(1/\gamma)+o(1)} \frac{\log(1/(\varepsilon\delta)) \log(1/\varepsilon)^4}{(1-\gamma)^5} \right) \quad \text{Setting } \eta = \gamma^{1/\max(2,\log(1/\varepsilon))} \\
&= \mathcal{O} \left((1/\varepsilon)^{\log(N\kappa)/\log(1/\gamma)+o(1)} \left(\frac{\log(1/\delta) + \log(1/\varepsilon)}{1-\gamma} \right)^5 \right)
\end{aligned}$$

If $N\kappa/\gamma^2 < 1$, then,

$$\begin{aligned}
T &\leq \mathcal{O} \left((1/\varepsilon)^2 \frac{1}{1-N\kappa/\gamma^2} \frac{\log(1/(\varepsilon\delta))}{(1-\eta)^2(1-\gamma)^2} \right) \\
&= \mathcal{O} \left((1/\varepsilon)^2 \frac{(\log(1/\delta) + \log(1/\varepsilon))^3}{(1-N\kappa/\gamma^2)(1-\gamma)^4} \right). \quad \text{Setting } \eta = \gamma^{1/\max(2,\log(1/\varepsilon))}
\end{aligned}$$

If $N\kappa/\gamma^2 = 1$, then,

$$\begin{aligned}
T &\leq \mathcal{O} \left(h_{\max}(1/\eta)^{h_{\max}} \frac{\log(1/(\varepsilon\delta))}{(1-\eta)^2(1-\gamma)^2\varepsilon^2} \right) \\
T &\leq \mathcal{O} \left(\frac{\log(1/(\varepsilon\delta)) \log(1/\varepsilon)^4}{(1-\gamma)^5\varepsilon^2} \right) \quad (1/\eta)^{h_{\max}} = \mathcal{O}(1) \\
T &\leq \mathcal{O} \left((1/\varepsilon)^2 \left(\frac{\log(1/\delta) + \log(1/\varepsilon)}{1-\gamma} \right)^5 \right).
\end{aligned}$$

□

C On the choice of the near-optimality definition

In this part, we discuss the possible alternative choices to the near-optimality set \mathcal{N}_h and explain the choice we made in Definition 1. From this definition, the knowledge of the rewards of the near-optimal nodes is enough to compute an optimal policy. Therefore, whatever the rewards associated with the non-near-optimal nodes are, they do not change the optimal policy. A good *adaptive* algorithm, would not exploring the whole tree, but would with high probability only explore a set not significantly larger than a set \mathcal{N}_h of near-optimal nodes. There are other possible definitions of the near-optimality set satisfying this desired property that we could possibly consider for the definition of \mathcal{N}_h . Ideally, this set would be as small as possible.

Alternative definition 1 An idea could be to consider for any node s and an ancestor s' , the value of the policy starting from s' choosing the action leading to s every time it can and playing optimally otherwise. Let $\mathcal{V}_s(s')$ be this value. A natural approach would be to define \mathcal{N}_h only based on $\mathcal{V}_s(s_0)$ with s_0 being the root. To ensure that exploring the near-optimal nodes is enough to compute the optimal policy we need any near-optimal nodes s of depth h to verify

$$(\mathcal{V}[s_0] - \mathcal{V}_s(s_0))p(s|s_0) \leq \frac{\gamma^h}{1-\gamma}.$$

When there is no **AVG** node, $p(s) = 1$ and this definition coincides with the near-optimality set defined in OLOP [5]. Nevertheless, in our case, $p(s)$ can be low and even 0 in the infinite case. When $p(s|s_0) = 0$ for all s , any node would be near-optimal which is bad because then the near-optimality set is large and the sample complexity of the resulting algorithm suffers.

Alternative definition 2 There is a smarter way to define a near-optimality set. We could define it as all the nodes s of depth h such that

$$\forall h' \leq h, \quad \sum_{i=h'}^h p(s_i|s_{h'}) \gamma^i (\mathcal{V}[s_i] - \mathcal{V}[c(s_i, s)]) \leq \frac{\gamma^h}{1-\gamma}.$$

Taking $h' = 0$, we recover the alternative definition 1. Alternative definition 2 defines a smaller near-optimality set, as the condition needs to be verified for all h' and not just for $h' = 0$. This definition would also lead to a smaller near-optimality set than our Definition 1. Indeed, in Definition 1 we consider only the first term of the sum above. This enables the algorithm to compute its strategy from a node s only based on the value of its descendants. This is an ingredient leading to better computational efficiency of `TrailBlazer`. Indeed, the computational complexity of our algorithm is linear in the number of calls to the generative model. In addition, when the transition probabilities are low, both definitions are close and coincide in the infinite case, when the probability transitions are continuous and therefore $p(s|s') = 0$.

D Auxiliary material

Lemma 4. *If for some $\alpha > 0, \beta, M$, a random variable X verifies that*

$$\mathbb{E}[e^{\lambda X}] \leq M \exp\left(\lambda|\alpha| + \frac{\beta^2 \lambda^2}{2}\right),$$

then for all $u > 0$

$$\mathbb{P}[|X| \geq u] \leq 2M \exp\left(-\frac{(u-\alpha)^2}{2\beta^2}\right).$$

Proof. Using Markov inequality for $u > \alpha$ and setting $\lambda = \frac{u-\alpha}{\beta^2}$, we get

$$\mathbb{P}[X \geq u] \leq \frac{\mathbb{E}[e^{\lambda X}]}{e^{\lambda u}} \leq M \exp\left(\lambda|\alpha| - \lambda u + \frac{\beta^2 \lambda^2}{2}\right) \leq M \exp\left(-\frac{(u-\alpha)^2}{2\beta^2}\right).$$

Similarly, for $u < -\alpha$,

$$\mathbb{P}[X \leq u] \leq M \exp\left(-\frac{(u+\alpha)^2}{2\beta^2}\right).$$

Taking a union bound, we have that for all $u > \alpha$,

$$\mathbb{P}[|X| \geq u] \leq 2M \exp\left(-\frac{(u-\alpha)^2}{2\beta^2}\right).$$

□

Lemma 5. *For all positive random variable X and event A*

$$\mathbb{E}[X|A] \leq \frac{\mathbb{E}[X]}{\mathbb{P}[A]}.$$

Proof.

$$\mathbb{E}[X] = \mathbb{E}[X|A] \mathbb{P}[A] + \mathbb{E}[X|A^c] \mathbb{P}[A^c] \geq \mathbb{E}[X|A] \mathbb{P}[A]$$

□