A VALUE OF COMPUTATION IN MCTS

If the state transition function is deterministic, then static and dynamic value computations simplify greatly:

$$\psi_n(s,a|\omega_{1:t}) = \mathbb{E}\left[\max_{(s',a')\in\Gamma_n(s)} Z(s',a'|\omega_{1:t})\right] \quad (3)$$

$$\phi_n(s,a|\omega_{1:t}) = \max_{(s',a')\in\Gamma_n(s)} \mathbb{E}\left[Z(s',a'|\omega_{1:t})\right] \ , \quad (4)$$

where $Z(s',a'|\omega_{1:t}) := r_1 + \gamma r_2 + \gamma^2 r_3 + \cdots + \gamma^n (Q_0^*(s',a')|\omega_{1:t})$, is the posterior leaf values scaled by γ^n and shifted by the discounted immediate rewards (r_i) along the path from s to s'.

In this 'flat' case, VOC(ϕ_n)-greedy policy is equivalent to a knowledge gradient policy, details of which can be found in Frazier et al. [5], Ryzhov et al. [19] for either isotropic or anisotropic Normal Q_0^*. On the other hand, VOC(ψ_n)-greedy policy has not been studied to the best of our knowledge. Computing the expected maximum of random variables is generally hard, which is required for ψ_n. Below, we offer a novel approximation to remedy this problem.

A.1 COMPUTING VOC(ψ_n)

We utilize a bound [12] that enables us to get a handle on ψ_n. This asserts,

$$\psi_n(s,a|\omega_{1:t}) \le c + \sum_{(s',a')\in\Gamma_n(s)} \int_c^\infty \left[1 - F_{s'a't}(x)\right] dx$$

for any $c \in \mathbb{R}$, where $F_{s'a't}$ is the CDF of $Z(s',a'|\omega_{1:t})$. This bound does not assume independence and holds for any correlation structure by assuming the worst case. Furthermore, the inequality is true for all c. However, the tightest bound is obtained by differentiating the RHS with respect to c, and setting its derivative to zero, which in turn yields $\sum_{(s',a')\in\Gamma_n(s)} \left[1 - F_{s'a't}(c)\right] = 1$ Thus, the optimizing c can be obtained via line search methods.

If $Z(\cdot,\cdot|\omega_{1:t})$ is distributed according to a multivariate (isotropic or anisotropic) Normal distribution, then we can eliminate the integral [15]:

$$\psi_n(s,a|\omega_{1:t}) \le \lambda_{sat} := c + \sum_{(s',a')\in\Gamma_n(s_\rho)} \Big[(\sigma_{s'a't})^2 F_{s'a't}(c) + (\mu_{s'a't} - c)[1 - F_{s'a't}(c)]\Big]$$

where $\mu_{s'a't}$ and $\sigma^2_{s'a't}$ are posterior mean and variances, that is $Z^*(s',a'|\omega_{1:t}) \sim \mathcal{N}(\mu_{s'a't}, (\sigma_{s'a't})^2)$.

If we further assume an isotropic Normal prior with mean $\mu_{s'a'0}$ and scale $\sigma_{s'a'0}$, and observation noise $\epsilon_i \sim \mathcal{N}(0,\sigma^2)$ i.i.d. for $i = 1, 2, \ldots, t$, then we get the posterior mean and scale as

$$\mu_{s'a't} = \frac{n_{s'a't}\hat{o}_{s'a't}/\sigma^2 + \mu_{s'a'0}/\sigma^2_{s'a't}}{n_{s'a't}/\sigma^2 + 1/\sigma^2_{s'a't}} \ ,$$
$$\sigma_{s'a't} = n_{s'a't}/\sigma^2_{s'a't} + 1/\sigma^2 \ ,$$

where $\hat{o}_{s'a't}$ is the mean trajectory rewards obtained from (s',a') and $n_{s'a't}$ is the number of times a sample is drawn from (s',a'). Then keeping c fixed, we can estimate the "sensitivity" of λ_{sat} with respect to an additional sample from (s',a') with

$$\frac{\partial \lambda_{sat}}{\partial n_{s'a't}} = \frac{\partial \lambda_{sat}}{\partial \sigma_{s'a't}} \frac{d\sigma_{s'a't}}{dn_{s'a't}} + \frac{\partial \lambda_{sat}}{\partial \mu_{s'a't}} \frac{d\mu_{s'a't}}{dn_{s'a't}} \ .$$

where

$$\frac{\partial \lambda_{sat}}{\partial \sigma_{s'a't}} = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-\left(\mu_{s'a't} - c\right)^2}{2\left(\sigma_{s'a't}\right)^2}\right)$$
$$\frac{d\sigma_{s'a't}}{dn_{s'a't}} = -\frac{\sigma\sigma^3_{sa0}}{2\left(n_{s'a't}\sigma^2_{sa0} + \sigma^2\right)^{\frac{3}{2}}}$$
$$\frac{\partial \lambda_{sat}}{\partial \mu_{s'a't}} = \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{\sqrt{2}\left(\mu_{s'a't} - c\right)}{2\sigma_{s'a't}}\right)\right)$$
$$\frac{d\mu_{s'a't}}{dn_{s'a't}} = \frac{\sigma^2\sigma^2_{sa0}\left(-\mu_{sa0} + \hat{o}_{s'a't}\right)}{(n_{s'a't})^2\sigma^4_{sa0} + 2n_{s'a't}\sigma^2\sigma^2_{sa0} + \sigma^4}.$$

We can then compute and utilize $\partial\lambda_{sat}/\partial n_{s'a't}$ as a proxy for the expected change in λ_{sat}. Because λ_{sat} is an upper bound, we find that this scheme works the best when the priors are optimistic, that is μ_{sa0} is large. In fact, as long as the prior mean is larger than the empirical mean, $\mu_{sa0} > \hat{o}_{s'a't}$, we have $\partial\lambda_{sat}/\partial n_{s'a't} < 0$. Then we can safely choose the best leaf to sample from via $\arg\min_{(s',a')\in\Gamma_n(s)} \left[\max_{a\in\mathcal{A}_s} \lambda_{sat}\right]$. We use this scheme when implementing VOC(ψ_n)-greedy in peg solitaire and confirmed that the results are nearly indistinguishable from calculating VOC(ψ_n)-greedy by drawing Monte Carlo samples in terms of the resulting regret curves.

B VOC-GREEDY ALGORITHMS

We provide the pseudocode for VOC-greedy MCTS policy in Algorithm 1. Throughout our analysis of this policy, we assume an infinite computation budget B.

B.1 TIME COMPLEXITIES

Computational complexity of VOC-greedy methods depend on a variety of factors, including the prior distribution of the leaf values, stochasticity, use of static vs

Algorithm 1: VOC(ϕ_n/ψ_n)-greedy for MCTS

Input: Current state s_ρ
Input: Maximum computation budget B
Output: Selected action to perform

Create a partial search graph/tree by expanding state s for n steps ;
Initialize the leaf set $\Gamma_n(s_\rho)$;
Initialize a partial function U, that maps states to UCT trees ;
$t \leftarrow 0$;
$\omega_{1:t} \leftarrow \epsilon$; `/* empty sequence */`
repeat
  $(s^*, a^*) = \arg\max_{\varpi} \text{VOC}_{\phi_n/\psi_n}(s_\rho, \varpi|\omega_{1:t})$;
  $s^\dagger \sim \mathcal{P}^{a^*}_{s^*\cdot}$;
  if $U(s^\dagger)$ *is not defined* **then**
    Initialize a UCT-tree rooted at $s^\dagger$;
    Define $U(s^\dagger)$, which maps to the UCT-tree from the previous step ;
  end
  Obtain sample $o_{s^*a^*t}$ by expanding $U(s^\dagger)$ and perform a roll-out ;
  $\omega_{t+1} \leftarrow (s^*, a^*, o_{s^*a^*t})$;
  $\omega_{1:t+1} \leftarrow \omega_{1:t}\omega_{t+1}$;
  $t \leftarrow t+1$;
until $\max_{\varpi}$ *VOC*$_{\phi_n/\psi_n}(s_\rho, \varpi|\omega_{1:t}) < 0$ ***or*** $t \geq B$;
return $\arg\max_{a \in \mathcal{A}_{s_\rho}} \phi_n/\psi_n(s_\rho, a|\omega_{1:t})$;

dynamic values. Here, we discuss the time complexity of computing the VOC(ϕ)-greedy policy with a conjugate Normal prior (with known variance) in MDPs with deterministic transitions.

The posterior values can be updated incrementally in constant time if the prior is isotropic Normal and in $O(m^2)$ if it is anisotropic, where m is the number of leaf nodes [3]. Given the posterior distributions, the value of a computation can be computed in $O(m)$ in the isotropic case and in $O(m^2 \log m)$ in the anisotropic case. We refer the reader to [5] for further details, as the analysis done for bandits with correlated Normal arms do apply directly.

C PROOFS

C.1 PROOF OF PROPOSITION 1

Let us consider the "base case" of $n = 1$ and define a higher-order function g, capturing the 1-step Bellman optimality equation for a state-action (s, a):

$$g(h) := \sum_{s'} \mathcal{P}^a_{ss'} \left[\mathcal{R}^a_{ss'} + \gamma \max_{a'} h(s', a') \right] .$$

Then $\phi_1(s, a|\omega_{1:t}) = g(\mathbb{E}[Q^*_0|\omega_{1:t}])$ and $\psi_1(s, a|\omega_{1:t}) = \mathbb{E}[g(Q^*_0|\omega_{1:t})]$. Because g is a convex function, we have $\psi_1(s, a|\omega_{1:t}) \geq \phi_1(s, a|\omega_{1:t})$ by Jensen's inequality. We use this to prove the upper bound in Proposition 1:

$$\begin{aligned} \psi_1(s, a|\omega_{1:t}) &= \mathbb{E}_{\Omega_{1:k}}[\psi_1(s, a|\omega_{1:t}\Omega_{1:k})] \\ &\geq \mathbb{E}_{\Omega_{1:k}}[\phi_1(s, a|\omega_{1:t}\Omega_{1:k})] , \end{aligned}$$

where the first inequality is due to Equation 1. For the lower bound, we have

$$\begin{aligned} \mathbb{E}_{\Omega_{1:k}}[\phi_1(s, a|\omega_{1:t}\Omega_{1:k})] &= \mathbb{E}_{\Omega_{1:k}}[g(\mathbb{E}[Q^*_0|\omega_{1:t}\Omega_{1:k}])] \\ &\geq g(\mathbb{E}_{\Omega_{1:k}}[\mathbb{E}[Q^*_0|\omega_{1:t}\Omega_{1:k})]] \\ &= g(\mathbb{E}[Q^*_0|\omega_{1:t}]) \\ &= \phi_1(s, a|\omega_{1:t}) . \end{aligned}$$

These inequalities also hold for $n > 1$ for the same reasons. We omit the proof.

C.2 PROOF OF PROPOSITION 2

Let ϖ^* denote the optimal candidate computation (of length 1), which minimizes Bayesian simple regret in expectation in one-step. That is,

$$\varpi^* := \arg\min_{\varpi} \mathbb{E}_\Omega[R_f(s_\rho, \omega_{1:t}\Omega)]$$

where $\Omega := (\varpi, O_{\varpi t+1})$ is the closure corresponding to ϖ. Then, we subtracting $R_f(s_\rho, \omega_{1:t})$, we get

$$\varpi^* = \arg\min_{\varpi} \left[\mathbb{E}_\Omega[R_f(s_\rho, \omega_{1:t}\Omega)] - R_f(s_\rho, \omega_{1:t})\right] .$$

The first terms of the regrets cancel out as $\mathbb{E}_{Q^*_0|\omega_{1:t}} \left[\max_{a \in \mathcal{A}_{s_\rho}} \Upsilon_n(s_\rho, a|\omega_{1:t})\right] = \mathbb{E}_\Omega \mathbb{E}_{Q^*_0|\omega_{1:t}\Omega} \left[\max_{a \in \mathcal{A}_{s_\rho}} \Upsilon_n(s_\rho, a|\omega_{1:t}\Omega)\right]$. Thus, we end up with,

$$\varpi^* = \arg\min_{\varpi} \left[-\mathbb{E}_\Omega \left[\max_a f(s_\rho, a|\omega_{1:t}\Omega) \right] + \max_a f(s_\rho, a|\omega_{1:t}) \right],$$

or equivalently $\varpi^* = \arg\max_{\varpi} \text{VOC}_f(s_\rho, \Omega|\omega_{1:t})$.

C.3 PROOF OF PROPOSITION 4

Consider the following 2-step (i.e., $n = 2$) search tree shown in Figure 6. The deterministic transitions are shown with the arrows, each corresponding to an action in $\mathcal{A} = \{\text{L}, \text{R}\}$. The leaves are denoted with filled circles whose posterior values are given by $Q^*_0(\cdot, \cdot)|\omega_{1:t}$. The root state is shown as s_ρ with its immediate successors as s_0 and s_1. Assume $\gamma = 1$ and all shown actions yield 0 immediate rewards. Finally, assume the posterior distribution of the leaf values are as in Figure 6, and they are pairwise independent.

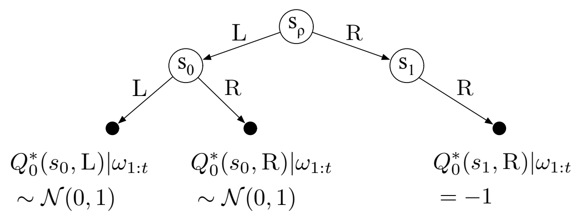

Figure 6: A search graph, where $\mathrm{VOC}'(\phi_n)$-greedy stops early.

In this case, we can see that no single sample from the leafs can result in a policy change at s_ρ since we would need to sample both of the leaves of the left subtree at least once for the policy at the root to change from L to R. Therefore, ${\mathrm{VOC}'}_{\phi_2}$ is zero for all possible computations here, and thus stops early, not achieving neither one-step nor asymptotic optimality. In contrast, VOC_{ϕ_2} is greater than zero for computations concerning the left subtree.

C.4 PROOF OF PROPOSITION 5

We need to show the equality of the second term in Equation 2 to the second term of VOC as we defined in Definition 3. First observe that $\mathbb{E}_{\Omega_{1:k}}[\psi_n(s_\rho, \alpha|\omega_{1:t}\Omega_{1:k})] = \psi_n(s_\rho, \alpha|\omega_{1:t})$. Then, we can take the α out as $\psi_n(s_\rho, \alpha|\omega_{1:t}) = \max_{a \in \mathcal{A}_{s_\rho}} \psi_n(s_\rho, a|\omega_{1:t})$, which is identical to the second term of our VOC definition in Equation 2.

D BANDIT TREE DETAILS

We utilizes trees of depth 7, where the agent transitions to the desired sub-tree with probability .75. In the correlated bandit arms case, the expected rewards of the arms are sampled from $\mathcal{N}(1/2, \Sigma)$ i.i.d. at each trial, where Σ is the covariance matrix given by an RBF kernel with scale parameter of 1 and the observation noise is sampled from $\mathcal{N}(0, 0.1)$ i.i.d. at each time step. In the uncorrelated case, the expected rewards are sampled from $\mathcal{U}(0.45, 0.55)$ and the observation noise is from $\mathcal{N}(0, 0.01)$. The former setting is designed to be noisier to compensate for the extra information provided by the correlations.