

# Variance-Aware Prior-Based Tree Policies for Monte Carlo Tree Search

Maximilian Weichart  
University of Regensburg

## Abstract

Monte Carlo Tree Search (MCTS) has profoundly influenced reinforcement learning (RL) by integrating planning and learning in tasks requiring long-horizon reasoning, exemplified by the AlphaZero family of algorithms. Central to MCTS is the search strategy, governed by a tree policy based on an upper confidence bound (UCB) applied to trees (UCT). A key factor in the success of AlphaZero is the introduction of a prior term in the *UCB1*-based tree policy *PUCT*, which improves exploration efficiency and thus accelerates training. While many alternative UCBs with stronger theoretical guarantees than *UCB1* exist, extending them to prior-based UCTs has been challenging, since *PUCT* was derived empirically rather than from first principles. Recent work retrospectively justified *PUCT* by framing MCTS as a regularized policy optimization (RPO) problem. Building on this perspective, we introduce *Inverse-RPO*, a general methodology that systematically derives prior-based UCTs from any prior-free UCB. Applying this method to the variance-aware *UCB-V*, we obtain two new prior-based tree policies that incorporate variance estimates into the search. Experiments indicate that these variance-aware prior-based UCTs outperform *PUCT* across multiple benchmarks without incurring additional computational cost. We also provide an extension of the `mctx` library supporting variance-aware UCTs, showing that the required code changes are minimal and intended to facilitate further research on principled prior-based UCTs. Code: <https://github.com/Max-We/inverse-rpo>.

## 1 Introduction

The combination of reinforcement learning (RL) with Monte Carlo Tree Search (MCTS) has led to major advances in artificial intelligence. Starting with AlphaGo (Silver et al., 2016), and subsequently generalized by AlphaZero (Silver et al., 2018) and MuZero (Schrittwieser et al., 2020), this line of work has achieved superhuman performance across domains requiring long-horizon reasoning and complex decision-making. These results underscore the power of integrating learning with search-based planning, and they motivate ongoing efforts to develop more efficient and broadly applicable variants of MCTS and AlphaZero-style methods.

A central component of MCTS is the tree policy, which balances exploration and exploitation to minimize regret. Before AlphaZero, such policies were derived from upper confidence bounds (UCBs) such as *UCB1* (Auer et al.), giving rise to the well-studied family of UCT algorithms, which apply UCBs to tree search. Over time, many variants beyond *UCB1*—including *UCB-V*, *Bayesian UCT*, and *UCB1-Uniform/Power* (Audibert et al., 2009; Tesauro et al., 2012; Asai and Wissow, 2024)—have been explored and shown to have a significant effect on the MCTS performance. With the AlphaZero family of algorithms, *UCB1* was extended by incorporating a prior term estimated by a neural network, yielding *PUCT*. This prior-based extension of *UCB1* greatly improved search efficiency in both small and large action spaces (Wu et al., 2023) and has since become the de facto standard tree policy. However, extending this prior-based approach to other UCBs has proven difficult. While the authors claim that *PUCT* is a variant of *PUCB* (Rosin, 2011), which in itself is an extension of *UCB1* with contextual information, a complete proof was never presented. Indeed, the concrete form of *PUCT* deviates from *UCB1* and *PUCB* by introducing a heuristic decay of the exploration term, and it is generally assumed to have been derived empirically rather than from formal guar-

antees<sup>1</sup>. We hypothesize that the extension of other UCBs to prior-based UCTs in the context of MCTS, although promising in theory, has been underexplored for that reason.

Table 1: Four prior-based UCT rules arranged by base UCB (columns) and heuristic form (rows). The heuristic form of the UCTs is described in Section 2.1. Our contributions are marked with \*.

	UCB1	UCB-V
<i>canonical form</i>	UCT-P	UCT-V-P*
<i>heuristic form</i>	PUCT	PUCT-V*

Recent work has reinterpreted MCTS as regularized policy optimization (RPO), showing that *PUCT* can be viewed as tracking the solution to a specific RPO. Our key insight is that this perspective not only provides an understanding for the form of prior-based UCBs in hindsight, such as previously described for *PUCT* (Grill et al., 2020), but also the theoretical foundation needed to systematically derive *any* prior-based UCT directly from prior-free UCBs by expressing them as an RPO. Building on this insight, we continue to study prior-based UCTs beyond *PUCT* by extending other, potentially stronger, UCB-based policies with prior terms. More concretely, we make the following key contributions:

**Inverse-RPO.** We introduce *Inverse-RPO*, a principled, step-by-step method that transforms a UCB into its prior-based counterpart. Unlike prior work that starts from an already prior-based selector such as *PUCT* (Grill et al., 2020) our method derives a prior-based selector systematically from its prior-free base form (e.g., *UCB1*). While prior work provides the formal framework linking MCTS and UCTs to RPO (Grill et al., 2020), we rearrange and slightly extend this approach into an easy-to-follow methodology, enabling researchers to apply it directly to their UCB of choice in future work.

**Variance-Aware Prior-Based UCTs.** To explore prior-based UCTs beyond *PUCT*, we instantiate *Inverse-RPO* on the variance-aware *UCB-V* to obtain two prior-based tree policies (see Table 1): (i) *UCT-V-P*, a principled RPO-derived variant; and (ii) *PUCT-V*, an heuristic analogue aligned with the practical form of *PUCT*. As experimental baselines, we compare these derived tree-policies against *PUCT* (the de facto choice in the AlphaZero family of algorithms), while also benchmarking against *UCT-P* (Grill et al., 2020), which can be viewed as a prior-based *UCB1* without the heuristic alterations introduced with *PUCT*.

<sup>1</sup>See the discussion by Grill et al. (2020) or the historical context in a Google Groups thread.

## Empirical Validation and Implementation.

Across a range of benchmark domains, we show that our variance-aware prior-based *UCT-V-P* and *PUCT-V* consistently match or outperform *UCT-P* and *PUCT* respectively, indicating that the benefits of replacing *UCB1* with stronger UCBs such as *UCB-V* in MCTS extend naturally to the prior-based MCTS as in the AlphaZero family of algorithms. We further propose an efficient implementation strategy for variance-aware MCTS, demonstrating that the derived *UCT-V-P* and *PUCT-V* can be deployed in practice as easily as the commonly used *PUCT* and at no extra computational overhead.

## 2 Preliminaries

Before presenting our methodology, we briefly review the key background concepts and notation needed throughout the paper. We begin with Monte Carlo Tree Search (MCTS) and its standard UCT formulation, followed by the regularized policy optimization (RPO) perspective that provides the foundation for our derivations.

### 2.1 Monte Carlo Tree Search

Monte Carlo Tree Search (MCTS) is a widely used planning algorithm that incrementally builds a search tree through repeated simulations (see Appendix B). During search, a tree policy based on an upper confidence bound (UCB) balances exploration and exploitation (Kocsis and Szepesvári, 2006)<sup>2</sup>. When *UCB1* is applied to trees, this yields the classical upper confidence bound for trees (*UCT1*) (Kocsis and Szepesvári, 2006):

$$\pi_{UCT1} \triangleq \arg \max_a \left[ q_a + c \cdot \sqrt{\frac{\log N}{1 + n_a}} \right]. \quad (1)$$

Here  $q_a$  is the empirical action value,  $n_a$  its visit count, and  $N = \sum_b n_b$  the total visits at the node. *UCT1* is provably optimal in the sense that it achieves the correct exploration-exploitation trade-off and converges to the optimal policy as the number of visits grows. Throughout this work, we add 1 to the visit count  $n_a$ , without loss of generality, to avoid division by zero and to simplify the subsequent analysis.

<sup>2</sup>Notation: (1) We use *UCB/UCT* in upright font as generic descriptors for the family of upper confidence bound rules (UCT denotes a UCB applied to trees). (2) Concrete algorithms/instantiations are written in italics (e.g., *UCB-V*, *PUCT*). (3) The canonical Hoeffding-based forms are written *UCB1/UCT1* to distinguish them from the generic descriptors in (1). A suffix “-P” indicates a prior-based extension (e.g., *UCT-P*, *PUCT-V*).

The action selection rule used in AlphaZero, commonly referred to as *PUCT* (Silver et al., 2017), was introduced later. It augments *UCB1* with the policy prior  $\pi_\theta(a)$ , which is being approximated by a neural network.

$$\pi_{\text{PUCT}} \triangleq \arg \max_a \left[ q_a + c \cdot \pi_\theta(a) \cdot \frac{\sqrt{N}}{1 + n_a} \right]. \quad (2)$$

**PUCT Heuristic Exploration Decay.** Besides the prior term, *PUCT* (2) departs from the principled *UCB1* rule by adopting a different exploration bonus that scales only with the square root of the total visit count  $N$ , rather than with  $\sqrt{\log N}$ . Formally, this amounts to replacing the exploration term

$$\sqrt{\frac{\log N}{1 + n_a}}$$

in *UCT1* (1) with

$$\frac{\sqrt{N}}{1 + n_a}.$$

Later, Grill et al. (2020) proposed a principled variant, *UCT-P*, which, similar to *PUCT* extends *UCB1* by incorporating the policy prior, but without the heuristic exploration decay.

$$\pi_{\text{UCT-P}} \triangleq \arg \max_a \left[ q_a + c \cdot \sqrt{\pi_\theta(a) \cdot \frac{\log N}{1 + n_a}} \right]. \quad (3)$$

By formalizing MCTS as a regularized policy optimization (RPO) problem, they showed that *UCT-P* directly expresses an RPO and that even *PUCT* can be cast within this framework—thus providing a theoretical justification in hindsight for its heuristic form.

## 2.2 Regularized Policy Optimization

Many machine-learning problems have been expressed as convex optimization problems (Bubeck, 2015), such as Support Vector Machines (SVMs) (Scholkopf and Smola) or Trust Region Policy Optimization (TRPO) (Schulman et al., 2017). Equivalently, reinforcement learning (RL) can be interpreted as a convex optimization problem by expressing it as RPO

$$\pi_{\theta'} \triangleq \arg \max_{\mathbf{y} \in \mathcal{S}} \left[ \mathbf{q}^\top \mathbf{y} - \mathcal{R}(\mathbf{y}, \pi_\theta) \right], \quad (4)$$

where  $\mathbf{y}$  is a distribution over actions,  $\mathbf{q}$  the corresponding  $q$ -values, and  $\mathcal{R} : \mathcal{S}^2 \rightarrow \mathbb{R}$  a divergence-based

convex regularizer that keeps  $\mathbf{y}$  close to the prior policy  $\pi_\theta$  (Neu et al., 2017; Geist et al., 2019; Grill et al., 2020).

Grill et al. (2020) proved that MCTS with *UCT1* (1) corresponds to the solution of an RPO with the Hellinger distance:

$$\begin{aligned} \bar{\pi}_{\text{UCT-P}} &\triangleq \arg \max_{\mathbf{y} \in \mathcal{S}} \left[ \mathbf{q}^\top \mathbf{y} - \lambda_N^{\text{UCT-P}} D_{\text{H}}(\pi_\theta, \mathbf{y}) \right], \\ \lambda_N^{\text{UCT-P}}(x) &\triangleq c \cdot \sqrt{\frac{\log N}{|\mathcal{A}| + N}}. \end{aligned} \quad (5)$$

where  $\mathcal{A}$  denotes the action set and  $\mathcal{S}$  is the  $|\mathcal{A}|$ -dimensional probability simplex.

Similarly, they showed that *PUCT* (2) expresses the solution to an RPO with the reverse-KL distance:

$$\begin{aligned} \bar{\pi}_{\text{PUCT}} &\triangleq \arg \max_{\mathbf{y} \in \mathcal{S}} \left[ \mathbf{q}^\top \mathbf{y} - \lambda_N^{\text{PUCT}} D_{\text{KL}}(\pi_\theta, \mathbf{y}) \right], \\ \lambda_N^{\text{PUCT}}(x) &\triangleq c \cdot \frac{\sqrt{N}}{|\mathcal{A}| + N}. \end{aligned} \quad (6)$$

From this RPO perspective, the *UCT-P* (3) and *PUCT* (2) can be recovered by considering the optimal action of the RPOs and evaluating the *marginal one-step gain* when selecting action  $a$ . Following prior work, we keep the notation  $\frac{\partial}{\partial n_a}$ ; operationally, this denotes the change along the coupled MCTS update in which both  $n_a$  and the total count  $N = \sum_b n_b$  increase by one.

$$a_{\text{UCT-P}}^* = \arg \max_a \left[ \frac{\partial}{\partial n_a} (\mathbf{q}^\top \hat{\pi} - \lambda_N^{\text{UCT-P}} D_{\text{H}}(\pi_\theta, \hat{\pi})) \right] \quad (7)$$

$$a_{\text{PUCT}}^* = \arg \max_a \left[ \frac{\partial}{\partial n_a} (\mathbf{q}^\top \hat{\pi} - \lambda_N^{\text{PUCT}} D_{\text{KL}}(\pi_\theta, \hat{\pi})) \right] \quad (8)$$

## 3 Deriving Prior-Based UCTs: Inverse RPO Pipeline

While previous work has established the existence of an RPO formulation for prior-based UCTs, a clear methodology to derive such a prior-based UCT starting from a prior-free UCB has been missing. Our first contribution is therefore *methodological*: we propose an *Inverse-RPO* pipeline, summarized in Figure 1, which provides a systematic procedure to derive prior-based UCTs from prior-free UCBs and offers researchers a principled framework to follow:

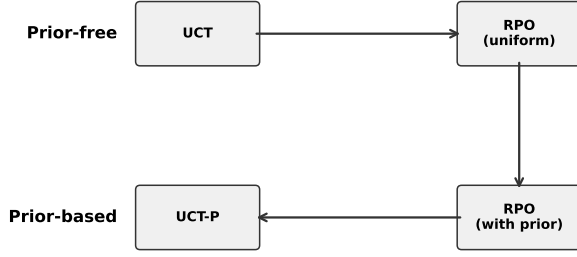


Figure 1: By casting the MCTS tree policy as the solution to an RPO objective, the prior becomes an explicit design term, yielding a principled prior-based UCT selection rule. This perspective resolves the otherwise opaque step from prior-free UCT to prior-based UCT and motivates our *Inverse-RPO* methodology.

1. **Factorize the UCT bonus.** Express the bonus in terms of the empirical visit distribution  $\hat{\pi}(a)$ , isolating a local term  $h(\hat{\pi}(a))$  from the global scaling factor  $\Phi(N)$ , i.e.  $B^{\text{UCT1}}(N, n_a) = \Phi(N) h(\hat{\pi}(a))$ .
2. **Define a separable  $f$ -regularizer.** Select a convex generator  $f$  such that  $f'(r) = -h(r)$ , yielding a prior-free RPO.
3. **Lift the regularizer with a prior.** Note that the prior-free RPO corresponds to the special case of an implicit prior-based RPO with uniform prior; generalize it by replacing the separable  $f$ -regularizer with a Csiszár  $f$ -divergence  $D_f(\pi_\theta, y)$ , thereby obtaining an explicit prior-based RPO.
4. **Recover the prior-based UCT rule.** Take the marginal gain with respect to  $n_a$  to derive the prior-based UCT selector.

### 3.1 UCT-P from UCT1: Applying the Inverse-RPO pipeline

For demonstration, we now apply the Inverse-RPO pipeline to the classical *UCT1* score (1) and obtain the prior-based rule *UCT-P* in (3). The same steps extend to other UCT-style scores (see Sec. 4 for *UCT-V*). Let a UCT-style selector  $S_a(q, n, N)$  and the empirical visit distribution  $\hat{\pi}(a)$  be defined as:

$$\begin{aligned} S_a(q, n, N) &= q_a + B(N, n_a), \\ N &= \sum_b n_b, \\ \hat{\pi}(a) &= \frac{1 + n_a}{|\mathcal{A}| + N}. \end{aligned} \quad (9)$$

Using this notation, the *UCT1* score (cf. Eq. (1)) becomes:

$$\begin{aligned} S_a^{\text{UCT1}}(q, n, N) &= q_a + B^{\text{UCT1}}(N, n_a), \\ B^{\text{UCT1}}(N, n_a) &= c \sqrt{\frac{\log N}{1 + n_a}}. \end{aligned} \quad (10)$$

#### 1. Factorize the UCT bonus.

We decompose the exploration term into a global scale  $\Phi(N)$  and a monotone shape function  $h$  of the empirical visit probability  $\hat{\pi}(a)$ . This separates the dependence on  $N$  and  $n_a$  and sets up the correspondence  $h = -f'$  used by the RPO derivation.

$$\begin{aligned} B^{\text{UCT1}}(N, n_a) &= \Phi(N) h(\hat{\pi}(a)), \\ h(r) &= \frac{1}{\sqrt{r}} \text{ (decreasing in } r), \\ \Phi(N) &= \frac{c \sqrt{\log N}}{\sqrt{|\mathcal{A}| + N}}. \end{aligned} \quad (11)$$

#### 2. Define a separable $f$ -regularizer.

Choose a convex generator whose (negative) derivative is  $h$ :

$$\begin{aligned} f^{\text{H}'}(r) &= -h(r) = -\frac{1}{\sqrt{r}}, \\ f^{\text{H}}(r) &= 2(1 - \sqrt{r}). \end{aligned} \quad (12)$$

In this case,  $f^{\text{H}}$  is the Hellinger function, which is convex and satisfies  $f^{\text{H}}(1) = 0$ . This yields the RPO with a *separable  $f$ -regularizer*:

$$\begin{aligned} L_{\text{UCT1}}(y) &= q^\top y - \lambda_N^{\text{UCT1}} \sum_a f^{\text{H}}(y_a), \\ \lambda_N^{\text{UCT1}} &= \Phi(N). \end{aligned} \quad (13)$$

Taking the marginal one-step gain with respect to  $n_a$  recovers the *UCT1* scoring rule matching (11):

$$\begin{aligned} a_{\text{UCT1}}^* &= \arg \max_a \frac{\partial}{\partial n_a} \left( q^\top \hat{\pi} - \lambda_N^{\text{UCT1}} \sum_b f^{\text{H}}(\hat{\pi}(b)) \right) \\ &= \arg \max_a \{ q_a + \Phi(N) h(\hat{\pi}(a)) \}. \end{aligned} \quad (14)$$

#### 3. Lift the regularizer with a prior.

We now *lift* the separable  $f$ -regularizer to the Csiszár  $f$ -divergence form with a prior  $\pi_\theta$ :

$$D_H(\pi_\theta, y) = \sum_a \pi_\theta(a) f^H\left(\frac{y_a}{\pi_\theta(a)}\right). \quad (15)$$

Utilizing the previously defined convex generator (12),  $D_H$  is a Hellinger-type  $f$ -divergence. Using this divergence, the prior-based RPO objective  $L_{\text{UCT-P}}$  and the corresponding greedy expansion rule  $a_{\text{UCT-P}}^*$  are identical to the ones presented by Grill et al. (2020):

$$L_{\text{UCT-P}}(y) = q^\top y - \lambda_N^{\text{UCT1}} D_H(\pi_\theta, y), \quad (16)$$

$$\lambda_N^{\text{UCT1}} = \Phi(N).$$

$$a_{\text{UCT-P}}^* = \arg \max_a \frac{\partial}{\partial n_a} \left( \mathbf{q}^\top \hat{\pi} - \lambda_N^{\text{UCT1}} D_H(\pi_\theta, \hat{\pi}) \right). \quad (17)$$

#### 4. Recover the prior-based UCT rule.

Solving the derivative condition in  $a_{\text{UCT-P}}^*$  and substituting  $f^H(r) = -h(r)$  yields the *UCT-P* selection rule. This rule coincides with the formulation of Grill et al. (2020) and can be interpreted as the prior-based analogue of the classical, prior-free *UCT* selection rule.

$$S_a^{\text{UCT-P}}(q, n, N) = q_a + \Phi(N) h\left(\frac{\hat{\pi}(a)}{\pi_\theta(a)}\right) \quad (18)$$

$$= q_a + c \sqrt{\pi_\theta(a) \cdot \frac{\log N}{1 + n_a}}.$$

#### 4 UCT-V-P and PUCT-V: Variance-Aware Prior-based UCTs

Our aim is to go beyond *UCB1*, studying alternative base UCBs with tighter confidence bonuses and deriving their prior-based counterparts via the *Inverse-RPO* pipeline. A natural candidate is *UCB-V*, which augments the exploration bonus with an empirical-variance term and is obtained from a Bernstein-type concentration inequality (in contrast to the Hoeffding inequality underlying *UCB1*) (Audibert et al., 2009). Under the same bounded-reward assumption, this yields variance-adaptive bonuses and correspondingly tighter instance-dependent guarantees than *UCB1*, without changing the problem setting. The variance-aware UCB-V applied to MCTS (Audibert et al., 2009; Wissow and Asai, 2024) is

$$S_a^{\text{UCT-V}}(q, n, N) = q_a + B^{\text{UCT-V}}(N, n_a, \hat{\sigma}_a^2), \quad (19)$$

$$B^{\text{UCT-V}}(N, n_a, \hat{\sigma}_a^2) \triangleq c_1 \hat{\sigma}_a \sqrt{\frac{\log N}{1 + n_a}} + c_2 \frac{\log N}{1 + n_a},$$

where  $\hat{\sigma}_a$  is the empirical reward standard deviation for action  $a$  consistent with earlier notation. We set  $c_1 = \sqrt{2}$  and  $c_2 = 3$ , so that the above expression is algebraically identical to the definition of Audibert et al. (2009), with the constants absorbed into  $c_1$  and  $c_2$ .

Analogous to the *PUCT* exploration-decay heuristic (see Section 2.1), we introduce an heuristic variant, *UCT-V-H*, which rewrites the exploration bonus as shown in (20). This heuristic form is introduced to make the comparison with *PUCT* meaningful as a whole; without it, we could only compare against the principled baseline *UCT-P*.

$$B^{\text{UCT-V-H}}(N, n_a, \hat{\sigma}_a^2) = c_1 \hat{\sigma}_a \frac{\sqrt{N}}{1 + n_a} + c_2 \frac{\log N}{1 + n_a}. \quad (20)$$

We apply the *Inverse-RPO* pipeline to obtain variance-aware, prior-based counterparts of *UCT-V* and its heuristic decay *UCT-V-H*. Specifically, the pipeline yields (i) UCT-style *selection rules* that can be used as drop-in replacements for *PUCT/UCT-P* during tree traversal and (ii) corresponding *RPO objectives* that mirror the selection rules in the optimization view of MCTS.

**Result: Variance-aware prior-based UCT selection rules.**

**UCT-V-P:**

$$S_a^{\text{UCT-V-P}}(q, n, N) = q_a + c_1 \cdot \hat{\sigma}_a \sqrt{\pi_\theta(a) \frac{\log N}{1 + n_a}} + c_2 \cdot \pi_\theta(a) \frac{\log N}{1 + n_a}. \quad (21)$$

**PUCT-V (heuristic prior-based variant):**

$$S_a^{\text{PUCT-V}}(q, n, N) = q_a + c_1 \cdot \pi_\theta(a) \hat{\sigma}_a \frac{\sqrt{N}}{1 + n_a} + c_2 \cdot \pi_\theta(a) \frac{\log N}{1 + n_a}. \quad (22)$$

*Derivations:* see Appendix C.

*Notable elements (selectors).* (i) The prior enters the exploration bonus as  $\pi_\theta(a)$ , reweighting both the variance and bias terms of *UCB-V*. (ii) The placement of the prior inside a square root for *UCT-V-P* follows from the divergences used in the Inverse-RPO lift (Hellinger vs. reverse-KL) and is specified in the next RPO objectives (other box). (iii) For a uniform prior, both selectors reduce to their prior-free forms.

**Result:** Variance-aware prior-based RPO targets.

**UCT-V-P:**

$$L_{\text{UCT-V-P}}(y) = \mathbf{q}^\top y - \lambda_N^{\text{UCT-V-1}} D_H(\pi_\theta, y) - \lambda_N^{\text{UCT-V-2}} D_{\text{KL}}(\pi_\theta, y), \quad (23)$$

$$\lambda_N^{\text{UCT-V-1}} = c_1 \frac{\sqrt{\log N}}{\sqrt{|\mathcal{A}| + N}}, \lambda_N^{\text{UCT-V-2}} = c_2 \frac{\log N}{|\mathcal{A}| + N}. \quad (24)$$

**PUCT-V (heuristic prior-based variant):**

$$L_{\text{PUCT-V}}(y) = \mathbf{q}^\top y - \lambda_N^{\text{UCT-V-H-1}} D_{\text{KL}}(\pi_\theta, y) - \lambda_N^{\text{UCT-V-H-2}} D_{\text{KL}}(\pi_\theta, y), \quad (25)$$

$$\lambda_N^{\text{UCT-V-H-1}} = c_1 \frac{\sqrt{N}}{|\mathcal{A}| + N}, \lambda_N^{\text{UCT-V-H-2}} = c_2 \frac{\log N}{|\mathcal{A}| + N}. \quad (26)$$

*Derivations:* see Appendix C.

*Notable elements (RPO objectives).* (i) In contrast to the *UCT-P* (5) and *PUCT* (6) optimization targets, which use a *single* regularizer term with one weight  $\lambda_N$ , our variance-aware contributions use *two* regularizer terms with distinct weights: a variance-term weight  $\lambda_N^{(1)}$  and a bias-term weight  $\lambda_N^{(2)}$ . (ii) As a result of the heuristic form of *UCT-V-H* in line with *PUCT*, the two variance-aware objectives are identical in their *second* regularizer term, and they differ only in the *first* regularizer and its weight

## 5 Experiments

Our experimental aim is twofold: (i) to implement the new variance-aware UCT policies *PUCT-V* and *UCT-V-P* introduced in Section 4; and (ii) to evaluate their performance relative to the classical prior-based baselines *PUCT* and *UCT-P*. We first describe the implementation details of the variance-aware extensions before turning to empirical comparisons.

### 5.1 Variance-aware MCTS Implementation

We provide a variance-aware MCTS implementation by extending the `mctx`<sup>3</sup> library (DeepMind et al., 2020). Enabling *UCT-V*-style rules requires propagating both empirical means and variances from a leaf to the root. To this end, we adopt Welford’s online update (see Algorithm 1), which is numerically stable and adds only a constant-time, constant-memory augmentation to the standard mean back-

propagation (Welford, 1962). Concretely, each node stores  $(n, \mu, \sigma^2)$  instead of  $(n, \mu)$ , where  $n$  is the visit count. The control flow and backward pass remain identical to standard mean backpropagation, with the starred (\*) lines denoting the added variance-tracking updates. During the selection phase, we also incorporate the proposed *PUCT-V* and *UCT-V-P* rules.

In the AlphaZero framework, a neural network is trained to approximate both the value function and the empirical visit distribution produced by MCTS. For our purposes, no additional variance head is required and the empirical variance from the tree search is sufficient.

**Algorithm 1** Variance-aware single-node update

**Input:** parent stats  $(n, \mu, \sigma^2)$ ; discounted value  $v = r + \gamma \cdot v_{\text{child}}$ .

**Complexity:** each update requires  $\mathcal{O}(1)$  arithmetic operations and  $\mathcal{O}(1)$  memory.

$$\begin{aligned} n^+ &\leftarrow n + 1, \\ \Delta &\leftarrow v - \mu, \\ \mu^+ &\leftarrow \mu + \frac{\Delta}{n^+}, \\ \Delta_2 &\leftarrow v - \mu^+ \quad * \\ \sigma^{2+} &\leftarrow \frac{n\sigma^2 + \Delta\Delta_2}{n^+} \quad * \end{aligned}$$

**Return:**  $(n^+, \mu^+, \sigma^{2+})$ .

Overall, adapting MCTS to be variance-aware and to use the proposed selection rules requires only three lines of code, excluding the additional variance field in the data structures.

### 5.2 Evaluation of PUCT-V and UCT-V-P

We evaluate on the **MinAtar** suite (Young and Tian, 2019), a widely used benchmark offering stochastic and deterministic Atari-style environments that preserve the core dynamics of the original games while being computationally efficient<sup>4</sup>. We access **MinAtar** through the **PGX** interface (Koyamada et al., 2023), which provides JAX-compatible environments and an open-source AlphaZero training script that we adapt for our experiments. The search/training pipeline is kept fixed across selectors to ensure a controlled comparison.

Unless otherwise noted, we run  $N_{\text{sim}} = 64$  simulations per move to generate training data. Evaluation is conducted at regular intervals in batches of 256 trajectories per seed, and with at least three seeds the per-

<sup>3</sup><https://github.com/google-deepmind/mctx>

<sup>4</sup>We exclude the *freeway* environment, as all evaluated algorithms consistently fail to achieve learning progress there.

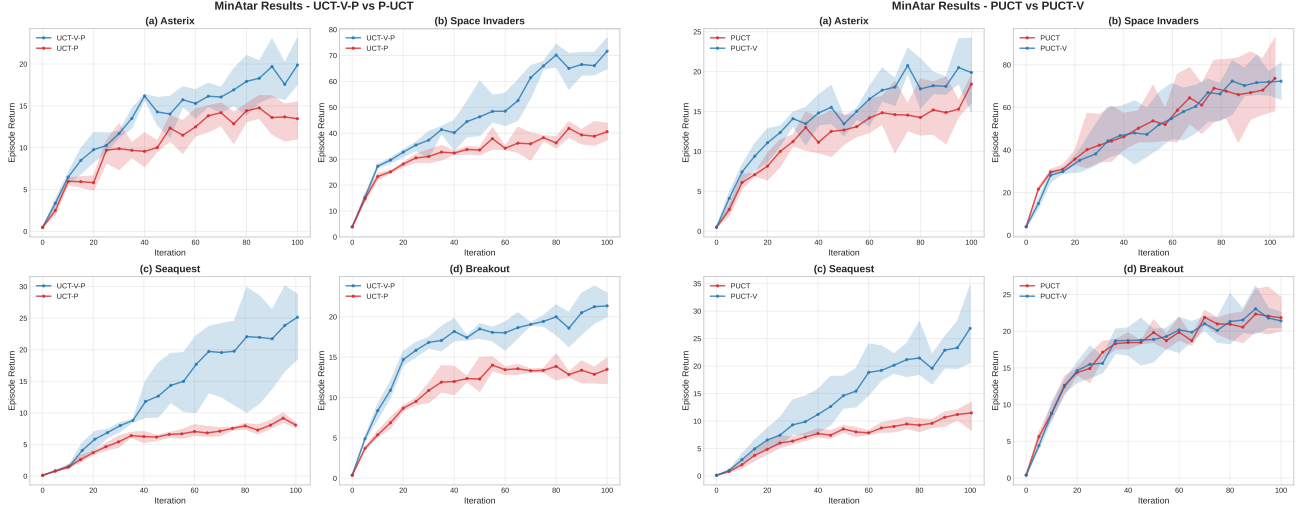


Figure 2: Average returns on the **MinAtar** suite with  $N_{\text{sim}} = 64$ . Evaluation is performed in batches of 256 per seed (at least 3 seeds), using only the trained policy head without MCTS. **Left:** *UCT-V-P* vs. *UCT-P*. **Right:** *PUCT* vs. *PUCT-V*. Solid lines indicate mean returns, and shaded regions show the corresponding best–worst range across seeds.

checkpoint estimates are sufficiently stable for meaningful comparisons. We adopt the network and optimization settings summarized in Table 2, holding hyperparameters constant across all methods to isolate the effect of the selection rule. Finally, we evaluate the learned *policy head* without search to assess representation and policy quality directly and to avoid confounding from test-time MCTS.

**Observations** Empirically, the measured wall-clock time per training step and per evaluation is essentially identical across selectors, indicating that the proposed variance-aware MCTS and selection rules incur no additional compute overhead. Figure 2 reports the average return of the trained policy head under all benchmarked selection rules. We compare *UCT-V-P* to *UCT-P* (heuristic-free) and *PUCT-V* to *PUCT* (heuristic-based). Across all environments, the variance-aware selectors match or exceed their variance-unaware baselines. In particular, *UCT-V-P* consistently outperforms *UCT-P*, showing that variance adjustment alone can substantially improve exploration. For the heuristic-based variants, *PUCT-V* surpasses *PUCT* on the stochastic games *Asterix* and *Seaquest*, and performs comparably on deterministic ones. Overall, variance-aware selection rules with priors yield consistent improvements, especially in stochastic settings, with negligible computational overhead and only minor modifications to MCTS.

## 6 Related Work

**AlphaZero family and prior-based tree policies.** Planning with MCTS coupled to learned function approximators became prominent with AlphaGo (Silver et al., 2016) and was iterated upon by AlphaZero (Silver et al., 2018) and MuZero (Schrittwieser et al., 2020). Furthermore, Stochastic MuZero (Antonoglou et al., 2022) handles stochastic dynamics while retaining *PUCT*, whereas Gumbel MuZero (Danihelka et al., 2022) adopts a Gumbel-based policy-improvement objective explicitly cast as regularized policy optimization (RPO). A unifying ingredient in these systems is a *prior-based* tree policy that injects a policy prior into the exploration bonus. Empirically, *PUCT* (and close relatives) has become the de facto choice across domains (Kemmerling et al., 2024).

**UCT family and stronger UCB bonuses.** Beyond *UCB1*, theoretically grounded UCT variants continue to be proposed (Browne et al., 2012). Among such developments, variance-aware Bernstein bonuses offer tighter instance-dependent guarantees under bounded rewards, which is why we select *UCB-V* (Audibert et al., 2009) as our base. Recent work explores alternative distributional assumptions (e.g., Gaussian and extreme-value regimes) with tailored regret analyses for classical planning (Wissow and Asai, 2024; Asai and Wissow, 2024). Notably, these methods are not prior-based by construction, so systematic prior-based extensions remain largely missing in the literature.

**Bayesian MCTS.** Variance-aware and uncertainty-quantifying approaches to MCTS are active research directions. Bayesian variants (*Bayes-UCT1/2*) maintain posteriors over node values and act via uncertainty bands (Tesauro et al., 2012); recent work explores richer uncertainty models and online inference (Greshler et al., 2024; Chen et al., 2025). While compelling, these methods typically introduce additional modelling choices, extra hyperparameters, and nontrivial bookkeeping. Our proposed variance-aware prior-based tree policies based on *UCB-V* likewise bring (frequentist) uncertainty quantification into the selection rule, yet integrate as drop-in replacements in the widely adopted AlphaZero-style MCTS with minimal changes.

**Regularized policy optimization (RPO) and MCTS.** Regularization-based views of RL connect policy improvement to convex programs with divergence penalties (Neu et al., 2017; Geist et al., 2019). Grill et al. (2020) brought this perspective to MCTS, thereby providing a retrospective theoretical understanding for prior-based tree policies such as *PUCT*. Follow-up analyses developed regret bounds for RPO-guided MCTS and studied entropy-based regularizers and backup operators (Dam et al., 2021). Complementing entropy-centric analyses, we focus on UCT-style bonuses by deriving variance-aware, prior-based selectors with matching RPO objectives (Eqs. 23 and 25).

## 7 Conclusion and Future Work

In this paper, we (1) proposed *Inverse-RPO*, a principled framework to derive prior-based UCTs from their prior-free base forms, and (2) instantiated this framework by deriving two prior-based versions of *UCB-V*. The resulting variance-aware prior-based tree-policies, *UCT-V-P* and *PUCT-V*, leverage variance estimates to improve search efficiency and outperform existing prior-based tree-policies *UCT-P* and *PUCT* across multiple benchmarks, with minimal implementation overhead.

Beyond the empirical results, our derivations of *UCT-V-P* and *PUCT-V* via the *Inverse-RPO* pipeline yield two RPO objectives that can be used as policy-training targets when casting MCTS as an optimization problem in future work. Another avenue for future work is to augment the network with a learned variance head, placed alongside the standard value and policy heads in the AlphaZero family, to refine search-based variance estimates and further improve the stability and performance of variance-aware prior-based UCTs. Finally, we invite the community to revisit the well-grounded UCB literature through this lens and make

principled use of its depth by systematically deriving yet underexplored prior-based UCTs.

## References

- 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, January 2016. ISSN 1476-4687. doi:10.1038/nature16961.
- David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharmashan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419):1140–1144, December 2018. doi:10.1126/science.aar6404.
- Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, December 2020. ISSN 1476-4687. doi:10.1038/s41586-020-03051-4.
- Peter Auer, Nicol O Cesa-Bianchi, and Paul Fischer. Finite-time Analysis of the Multiarmed Bandit Problem.
- Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration–exploitation trade-off using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, April 2009. ISSN 0304-3975. doi:10.1016/j.tcs.2009.01.016.
- Gerald Tesauro, V. T. Rajan, and Richard Segal. Bayesian Inference in Monte-Carlo Tree Search, March 2012.
- Masataro Asai and Stephen Wissow. Extreme Value Monte Carlo Tree Search, May 2024.
- Ti-Rong Wu, Hung Guei, Pei-Chiun Peng, Po-Wei Huang, Ting Han Wei, Chung-Chin Shih, and Yun-Jui Tsai. MiniZero: Comparative Analysis of AlphaZero and MuZero on Go, Othello, and Atari Games. <https://arxiv.org/abs/2310.11305v3>, October 2023.



- Christopher D. Rosin. Multi-armed bandits with episode context. *Annals of Mathematics and Artificial Intelligence*, 61(3):203–230, March 2011. ISSN 1012-2443, 1573-7470. doi:10.1007/s10472-011-9258-6.
- Jean-Bastien Grill, Florent Altché, Yunhao Tang, Thomas Hubert, Michal Valko, Ioannis Antonoglou, and Rémi Munos. Monte-Carlo Tree Search as Regularized Policy Optimization, July 2020.
- Levente Kocsis and Csaba Szepesvári. Bandit Based Monte-Carlo Planning. In Johannes Fürnkranz, Tobias Scheffer, and Myra Spiliopoulou, editors, *Machine Learning: ECML 2006*, pages 282–293, Berlin, Heidelberg, 2006. Springer. ISBN 978-3-540-46056-5. doi:10.1007/11871842\_29.
- David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dhharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. AlphaZero Chess - Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm, December 2017.
- Sébastien Bubeck. Convex Optimization: Algorithms and Complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015. ISSN 1935-8237, 1935-8245. doi:10.1561/22000000050.
- Bernhard Scholkopf and Alex Smola. Support Vector Machines and Kernel Algorithms.
- John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust Region Policy Optimization, April 2017.
- Gergely Neu, Anders Jonsson, and Vicenç Gómez. A unified view of entropy-regularized Markov decision processes, May 2017.
- Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A Theory of Regularized Markov Decision Processes, June 2019.
- Stephen Wissow and Masataro Asai. Scale-Adaptive Balancing of Exploration and Exploitation in Classical Planning, August 2024.
- DeepMind, Igor Babuschkin, Kate Baumli, Alison Bell, Surya Bhupatiraju, Jake Bruce, Peter Buchlovsky, David Budden, Trevor Cai, Aidan Clark, Ivo Danihelka, Antoine Dedieu, Claudio Fantacci, Jonathan Godwin, Chris Jones, Ross Hemsley, Tom Hennigan, Matteo Hessel, Shaobo Hou, Steven Kapturowski, Thomas Keck, Iurii Kemaev, Michael King, Markus Kunesch, Lena Martens, Hamza Merzic, Vladimir Mikulik, Tamara Norman, George Papamakarios, John Quan, Roman Ring, Francisco Ruiz, Alvaro Sanchez, Laurent Sartran, Rosalia Schneider, Eren Sezener, Stephen Spencer, Srivatsan Srinivasan, Miloš Stanojević, Wojciech Stokowiec, Luyu Wang, Guangyao Zhou, and Fabio Viola. The DeepMind JAX Ecosystem, 2020.
- B. P. Welford. Note on a Method for Calculating Corrected Sums of Squares and Products. *Technometrics*, 4(3):419–420, 1962. ISSN 00401706. doi:10.2307/1266577.
- Kenny Young and Tian Tian. MinAtar: An Atari-Inspired Testbed for Thorough and Reproducible Reinforcement Learning Experiments, June 2019.
- Sotetsu Koyamada, Shinri Okano, Soichiro Nishimori, Yu Murata, Keigo Habara, Haruka Kita, and Shin Ishii. Pgx: Hardware-Accelerated Parallel Game Simulators for Reinforcement Learning. *Advances in Neural Information Processing Systems*, 36:45716–45743, December 2023.
- Ioannis Antonoglou, Julian Schrittwieser, Sherril Ozair, and Thomas Hubert. PLANNING IN STOCHASTIC ENVIRONMENTS WITH A LEARNED MODEL. 2022.
- Ivo Danihelka, Arthur Guez, Julian Schrittwieser, and David Silver. POLICY IMPROVEMENT BY PLANNING WITH GUMBEL. 2022.
- Marco Kemmerling, Daniel Lütticke, and Robert H. Schmitt. Beyond Games: A Systematic Review of Neural Monte Carlo Tree Search Applications. *Applied Intelligence*, 54(1):1020–1046, January 2024. ISSN 0924-669X, 1573-7497. doi:10.1007/s10489-023-05240-w.
- 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, March 2012. ISSN 1943-068X, 1943-0698. doi:10.1109/TCIAIG.2012.2186810.
- Nir Greshler, David Ben Eli, Carmel Rabinovitz, Gabi Guetta, Liran Gispán, Guy Zohar, and Aviv Tamar. A Bayesian Approach to Online Planning, June 2024.
- Jiayu Chen, Wentse Chen, and Jeff Schneider. Bayes Adaptive Monte Carlo Tree Search for Offline Model-based Reinforcement Learning, May 2025.
- Tuan Q. Dam, Carlo D’Eramo, Jan Peters, and Joni Pajarinen. Convex Regularization in Monte-Carlo Tree Search. In *Proceedings of the 38th International Conference on Machine Learning*, pages 2365–2375. PMLR, July 2021.

---

# Variance-Aware Prior-Based Tree Policies for Monte Carlo Tree Search:

## Supplementary Materials

---

### A Supplementary Code Release

The full source code and reproduction instructions are available at: [github.com/Max-We/inverse-rpo](https://github.com/Max-We/inverse-rpo).

- **Modifications to `mctx`.** This includes the variance-aware extensions of the MCTS backpropagation routine, together with the implementation of the proposed variance-aware tree policies *UCT-V-P* and *PUCT-V*. These modifications are fully integrated into the existing `mctx` API and intended as minimal drop-in changes.
- **Training and Evaluation.** We include the `pgx` environments (Koyamada et al., 2023) with a training script adapted for the `MinAtar` experiments. This script reproduces all experimental results presented in Section 5.

### B Monte Carlo Tree Search: Four Stages

For completeness, we briefly recall the four canonical stages of MCTS:

1. **Selection.** Starting from the root, recursively select child nodes according to a tree policy (e.g., *UCT1* or *PUCT*) until reaching a leaf node.
2. **Expansion.** If the leaf node corresponds to a non-terminal state, expand the tree by adding a child node to the selected leaf node. Some implementations also expand terminal nodes, setting the discount-factor  $\gamma$  to zero.
3. **Evaluation (Simulation).** Evaluate the expanded node with a neural network (AlphaZero approach) or by conducting rollouts under a rollout policy (classical approach).
4. **Backpropagation.** Propagate the evaluated node statistics back through the visited path, updating statistics at each parent node. These statistics typically include the information required by the tree policy, such as the value and visit count of a node.

These four steps are repeated for a fixed number of simulations, after which an action is chosen based on the statistics of the children of the root node. In the AlphaZero family of algorithms, a neural network is additionally trained on the root statistics and node statistics such as the value are **normalized** to conform to the UCB assumptions.

### C Inverse-RPO Derivations for Variance-Aware UCTs

This appendix provides the complete inverse-RPO derivations that lead to the prior-based, variance-aware selection rules and objectives stated in the main text (§4; cf. (21), (22), (23), (25)). We reuse the empirical selector  $\hat{\pi}$  from (9).

#### C.1 UCT-V-P (derivation)

1. **Factorize the UCT bonus.** Starting from the variance-aware UCT score (19), the exploration bonus factorizes as

$$B^{\text{UCT-V}}(N, n_a, \hat{\sigma}_a^2) = \lambda_N^{\text{UCT-V}-1} h_{\text{H}}(\hat{\pi}(a), \hat{\sigma}_a) + \lambda_N^{\text{UCT-V}-2} h_{\text{KL}}(\hat{\pi}(a)), \quad h_{\text{H}}(r, \sigma) = \frac{\sigma}{\sqrt{r}}, \quad h_{\text{KL}}(r) = \frac{1}{r}, \quad (27)$$

with scaling terms

$$\lambda_N^{\text{UCT-V-1}} = c_1 \frac{\sqrt{\log N}}{\sqrt{|\mathcal{A}|+N}}, \quad \lambda_N^{\text{UCT-V-2}} = c_2 \frac{\log N}{|\mathcal{A}|+N}. \quad (28)$$

**2. Define a separable  $f$ -regularizer.** Choose convex generators whose (negative) derivatives match  $h_H$  and  $h_{KL}$ :

$$f^H(r, \sigma) = 2\sigma(1 - \sqrt{r}) \Rightarrow f^{H'}(r, \sigma) = -\frac{\sigma}{\sqrt{r}}, \quad f^{KL}(r) = -\log r \Rightarrow f^{KL'}(r) = -\frac{1}{r}. \quad (29)$$

This yields the RPO with a *separable  $f$ -regularizer*:

$$L_{\text{UCT-V}}(y) = \mathbf{q}^\top y - \lambda_N^{\text{UCT-V-1}} \sum_a f^H(y_a, \hat{\sigma}_a) - \lambda_N^{\text{UCT-V-2}} \sum_a f^{KL}(y_a), \quad (30)$$

whose marginal-gain rule in  $n_a$  recovers (27).

**3. Lift the regularizer with a prior.** Lift the separable  $f$ -regularizers to (weighted) Csiszár forms with prior  $\pi_\theta$ :

$$D_H(\pi_\theta, y) = \sum_a \pi_\theta(a) f^H\left(\frac{y_a}{\pi_\theta(a)}, \hat{\sigma}_a\right), \quad D_{KL}(\pi_\theta, y) = \sum_a \pi_\theta(a) f^{KL}\left(\frac{y_a}{\pi_\theta(a)}\right). \quad (31)$$

The prior-based objective is exactly the form stated in the main text:

$$\text{(cf. (23))} \quad L_{\text{UCT-V-P}}(y) = \mathbf{q}^\top y - \lambda_N^{\text{UCT-V-1}} D_H(\pi_\theta, y) - \lambda_N^{\text{UCT-V-2}} D_{KL}(\pi_\theta, y).$$

**4. Recover the prior-based UCT rule.** Taking the directional derivative in  $n_a$  yields the greedy expansion rule reported in the main text:

$$\text{(cf. (21))} \quad S_a^{\text{UCT-V-P}}(q, n, N) = q_a + c_1 \cdot \hat{\sigma}_a \sqrt{\pi_\theta(a) \frac{\log N}{1+n_a}} + c_2 \cdot \pi_\theta(a) \frac{\log N}{1+n_a}.$$

## C.2 PUCT-V (heuristic variant; derivation)

**1. Factorize the UCT bonus.** For the heuristic variant (20), the bonus factorizes as

$$B^{\text{UCT-V-H}}(N, n_a, \hat{\sigma}_a^2) = \lambda_N^{\text{UCT-V-H-1}} h_H(\hat{\pi}(a), \hat{\sigma}_a) + \lambda_N^{\text{UCT-V-H-2}} h_{KL}(\hat{\pi}(a)), \quad (32)$$

with

$$h_H(r, \sigma) = \frac{\sigma}{r}, \quad h_{KL}(r) = \frac{1}{r}, \quad \lambda_N^{\text{UCT-V-H-1}} = c_1 \frac{\sqrt{N}}{|\mathcal{A}|+N}, \quad \lambda_N^{\text{UCT-V-H-2}} = c_2 \frac{\log N}{|\mathcal{A}|+N}. \quad (33)$$

**2. Define a separable  $f$ -regularizer.** Choose convex generators with (negative) derivatives  $h_H$  and  $h_{KL}$ :

$$f^{KL}(r, \sigma) = -\sigma \log r \Rightarrow f^{KL'}(r, \sigma) = -\frac{\sigma}{r}, \quad f^{KL}(r) = -\log r \Rightarrow f^{KL'}(r) = -\frac{1}{r}. \quad (34)$$

This yields the RPO with a *separable  $f$ -regularizer*:

$$L_{\text{UCT-V-H}}(y) = \mathbf{q}^\top y - \lambda_N^{\text{UCT-V-H-1}} \sum_a f^{KL}(y_a, \hat{\sigma}_a) - \lambda_N^{\text{UCT-V-H-2}} \sum_a f^{KL}(y_a), \quad (35)$$

whose marginal-gain rule recovers (32).

**3. Lift the regularizer with a prior.** Lifting to Csiszár forms with prior  $\pi_\theta$  gives the prior-based objective reported in the main text:

$$\text{(cf. (25))} \quad L_{\text{PUCT-V}}(y) = \mathbf{q}^\top y - \lambda_N^{\text{UCT-V-H-1}} D_{KL}(\pi_\theta, y) - \lambda_N^{\text{UCT-V-H-2}} D_{KL}(\pi_\theta, y).$$

**4. Recover the prior-based UCT rule.** Taking the directional derivative in  $n_a$  yields the selection rule as stated:

$$\text{(cf. (22))} \quad S_a^{\text{PUCT-V}}(q, n, N) = q_a + c_1 \cdot \pi_\theta(a) \hat{\sigma}_a \frac{\sqrt{N}}{1+n_a} + c_2 \cdot \pi_\theta(a) \frac{\log N}{1+n_a}.$$

## D Hyperparameters for the MinAtar Experiments

In our experiments we used the hyperparameters in Table 2 consistently across all benchmarks.

Table 2: Key hyperparameters used for MinAtar experiments.

<b>Hyperparameters</b>	<b>Value</b>
Iterations	100
Simulations ( $N_{\text{sim}}$ )	64
Self-play batch size	256
Max. steps per episode	256
Training batch size	1024
Learning rate	$1 \times 10^{-3}$
Discount factor $\gamma$	0.99
Evaluation interval	every 5 iterations
Network architecture	6-layer ResNet-v2, 32 channels