Workflow Search Reinforcement Learning over Structured Decompositions

Guangyu Jiang GUANGYU.JIANG@GWU.EDU
Department of Electrical and Computer Engineering, George Washington University

Shu Hong SHU.HONG@GWU.EDU
Department of Electrical and Computer Engineering, George Washington University

Mahdi Imani M.IMANI@NORTHEASTERN.EDU
Department of Electrical and Computer Engineering, Northeastern University

Nathaniel D. Bastian NATHANIEL.BASTIAN@WESTPOINT.EDU
Department of Electrical Engineering and Computer Science, United States Military Academy

Tian Lan TLAN@GWU.EDU
Department of Electrical and Computer Engineering, George Washington University

Editors: G. Sukhatme, L. Lindemann, S. Tu, A. Wierman, N. Atanasov

Abstract

We study workflow search reinforcement learning (RL) for long-horizon tasks that can be decomposed into ordered, semantically interpretable subtasks. A workflow specifies an ordered set of milestones or procedural steps. Rather than learning a library of low-level skills and a meta-controller, we treat the set of feasible workflows as the high-level search domain. We then train a workflow-conditioned policy in an inner reinforcement learning loop. We propose a Gaussian process upper confidence bound workflow search (GP-UCB-WS) method. It places a Gaussian process prior over the workflow-to-return map and uses the upper confidence bound rule to adaptively select promising workflows. For each selected workflow, a base RL algorithm optimizes the corresponding conditioned policy using a shaped reward. We derive regret bounds that decompose the overall error into (i) Bayesian optimization error in workflow space and (ii) a policy-learning error for the workflow-conditioned inner loop, yielding provable regret bounds with respect to the optimal workflow and policy. In compositional tasks, including an ordered-visit gridworld and the TTCP CAGE Challenge 2 cyber defense environment, GP-UCB-WS significantly accelerates learning and achieves higher or comparable returns than flat proximal policy optimization (PPO), soft actor critic (SAC), and hierarchical RL (HRL) baselines, particularly when the workflow representation captures latent low-dimensional structure of the learning problems.

Keywords: Reinforcement Learning, Bayesian Optimization, Workflow Search.

1. Introduction

Long-horizon decision-making often decomposes into subtasks whose ordering determines success, such as collecting parts before assembly or unlocking rooms before navigation. In these settings, performance depends not only on competent low-level control, but also on choosing an effective ordering of milestones. We therefore view the order itself as a decision variable that should be explored and scored during learning.

Conventional hierarchical reinforcement learning (HRL) tackles long-horizon problems by learning a high-level controller that selects among primitive or temporally extended low-level policies (Sutton et al., 1999; Dietterich, 2000). In practice, the meta-controller is trained via trial and error,

which tends to be data-hungry and often neglects the available domain structure (Levy et al., 2017; Bacon et al., 2017; Verma et al., 2018; Nachum et al., 2018). More recent hierarchical methods incorporate planning priors or predictive models to mitigate exploration costs. However, they still require substantial search effort to uncover effective decompositions (Tamar et al., 2016; Farquhar et al., 2017; Racanière et al., 2017; Yu, 2025). At the other extreme, one-shot planners, whether symbolic pipelines or large language model (LLM)-generated scripts, produce fixed high-level sequences that rarely adapt dynamically through feedback. Rule- or ontology-driven systems encode precedence relations but hard-code the ordering of subtasks, which can degrade performance when the order is suboptimal or task conditions change (Zhuang et al.).

In this work, we take the view that many long-horizon tasks admit multiple valid decompositions and that the choice of decomposition should itself be explored. Rather than committing to a single plan or learning a meta-policy from scratch, we treat the workflow (i.e., an ordered sequence of milestones) as an explicit high-level decision variable. Our algorithm, GP-UCB-WS, actively explores and scores candidate workflows while training a policy conditioned on the selected workflow, thus decoupling the discrete ordering problem from continuous control. Concretely, we begin with a set of candidate subtasks or milestones (e.g., derived from domain expertise or LLM-based extraction) and define a finite set of feasible workflows over them. At each outer iteration, the algorithm selects a workflow using a Gaussian process (GP) surrogate over workflow embeddings with an upper confidence bound (UCB) rule, then runs a fixed-budget inner loop that trains a workflow-conditioned policy with potential-based reward shaping to promote adherence to the prescribed order. Crucially, policy evaluation is always performed under the original environment reward, so the outer objective remains the same as in the unshaped MDP. The kernel on workflows enables generalization across similar orders, allowing the GP surrogate to infer promising candidates without retraining every permutation from scratch, and thereby enabling data-efficient search in a combinatorial space.

This design leverages partial task structure without fixing a single plan, focuses computation on promising workflows instead of treating the combinatorial space as flat, and produces interpretable outer-loop traces. Theoretically, we obtain end-to-end regret bounds that decompose into Bayesian optimization regret over workflows, an inner fixed-budget learning gap, and a best-of-M selection-bias term. Empirically, on an ordered-visit gridworld and TTCP CAGE Challenge 2, GP-UCB-WS matches or exceeds flat PPO/SAC and HRL baselines, with the largest gains when the workflow embedding reflects underlying task structure.

Our contributions are as follows:

- **Formulation.** We formalize workflow search RL as a bi-level problem: an outer search over a finite, kernelized workflow set and an inner learner that optimizes a workflow-conditioned policy under a fixed budget.
- **Algorithm.** We introduce GP-UCB-WS: a GP surrogate on workflow embeddings with UCB sampling, coupled with PPO training that uses potential-based shaping for faster learning and evaluates with the original reward to preserve the outer objective.
- **Theory.** We derive end-to-end regret bounds that decompose into (i) Bayesian optimization regret over workflows, (ii) an inner finite-time policy-learning gap, and (iii) a small selection-bias term from using best-of-M training means without independent evaluation. Anchored PBRS preserves the outer objective, and the noise model supports standard finite-domain GP-UCB guarantees.

- **Empiricals.** On an ordered-visit gridworld and the TTCP CAGE Challenge 2, GP-UCB-WS achieves faster learning and comparable or higher final returns than flat PPO/SAC and hierarchical RL baselines, especially when the workflow embedding reflects a low-dimensional structure. The outer loop leaves an interpretable trace of selected workflows, revealing how the agent explores and refines task decompositions over time.

2. Related Work

Hierarchical reinforcement learning. In RL, hierarchical methods have addressed long-horizon tasks by decomposing decision-making into multiple levels of abstraction. Classic architectures such as FeUdal Networks (Vezhnevets et al., 2017), Option-Critic (Bacon et al., 2017), and HIRO (Nachum et al., 2018) learn high-level policies that select among low-level skills or options, enabling agents to plan over extended time spans with fewer decision points. However, these approaches often treat both levels under the same learning paradigm and require substantial sample budgets, frequently failing to exploit domain structures or semantic guidance. Complementary unsupervised skill discovery methods such as DIAYN (Eysenbach et al.), Variational Intrinsic Control (Gregor et al., 2016), DADS (Sharma et al.), VISR (Hansen et al.), and Contrastive Intrinsic Control (Laskin et al., 2022) utilize mutual information or state coverage objectives to discover diverse behaviors. However, they often produce latent skills that diverge from semantically meaningful subtasks and struggle in environments with structured dependencies. Extensions like Lipschitz-constrained skill discovery (Park et al.) or controllability-aware discovery (Park et al., 2023) prioritize spatial coverage or controllability but still do not capture task-specific sequence structure.

Symbolic and programmatic task decompositions. A different thread of work conditions neural policies on symbolic plans or sketches. The policy sketches framework associates tasks with sequences of abstract subtasks and trains modular subpolicies to align with those sketches, enabling reuse across tasks without low-level supervision (Andreas et al., 2017). However, the sketch is fixed ahead of time, and the method does not explore alternative decompositions. Relatedly, reward machines or automata-style decompositions embed symbolic task logic; for example, they decompose non-Markovian reward functions into automaton states and guide modular policy learning (Toro Icarte et al., 2019; Icarte et al., 2022). These methods typically require that the automaton structure is given or inferred separately and do not integrate adaptive planning over decompositions within the main learning loop. Programmatic RL methods, such as PIRL (Verma et al., 2018) or more general logic-augmented policy synthesis, constrain policies using grammars or domain-specific languages, and blend planning and execution into unified programmatic controllers. However, they do not clearly separate high-level exploration of alternative decompositions from low-level policy refinement in the manner of our workflow search approach.

Bayesian optimization for structured spaces. In the domain of Bayesian optimization (BO), the classical BO applies to continuous or low-dimensional inputs. More recent work has extended BO to structured, combinatorial, or graph-structured input spaces. COMBO (Oh et al., 2019) introduces diffusion kernels over discrete variable graphs to allow GP over combinatorial input domains. Other works, such as LADDER, merge latent space embeddings with structured kernels to better model black-box functions on sequences or graphs (Deshwal and Doppa, 2021). Although BO has found use in RL for hyperparameter tuning and architecture search, it has not been leveraged to explore workflow or decomposition spaces that directly shape agent behavior.

In summary, prior work either commits to a high-level structure a priori or learns all levels without a mechanism to compare alternative decompositions; unsupervised methods add skill diversity but rarely yield task-aligned structure. We instead treat an ordered decomposition as a first-class decision variable and place a Bayesian optimization loop over this discrete, structured space. Concretely, a GP surrogate with UCB acquisition proposes candidate workflows, while an inner loop trains workflow-conditioned policies and returns performance as feedback. This separation provides an uncertainty-aware and sample-efficient way to explore plausible decompositions while retaining interpretability and flexibility for long-horizon RL.

3. Problem Formulation

We consider a Markov Decision Process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma, \rho_0)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P(s' \mid s, a)$ is the transition function, $r(s, a, s')$ is the reward function, $\gamma \in (0, 1)$ is the discount factor, and ρ_0 is the initial state distribution. Let $\mathcal{W}$ denote a set of feasible workflows. Each $w \in \mathcal{W}$ encodes a structured decomposition of the task (e.g. an ordered sequence of subtasks, possibly with parameters). We assume an embedding $\phi : \mathcal{W} \to \mathbb{R}^d$ and a positive semidefinite kernel $k_{\mathbb{R}^d} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. This induces a kernel on workflows $k_{\mathcal{W}}(w, w') := k_{\mathbb{R}^d}\big(\phi(w), \phi(w')\big)$, which quantifies structural similarity between workflow candidates.

Given a workflow w, we define a workflow-conditioned policy $\pi(\cdot \mid s, w)$ that takes workflow w as input and follows the task decomposition specified by the workflow w. The expected discounted return is provided in (1).

$$J_r(\pi; w) = \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t \, r(s_t, a_t, s_{t+1}) \;\Big|\; a_t \sim \pi(\cdot \mid s_t, w),\; s_{t+1} \sim P(\cdot \mid s_t, a_t)\Big]. \tag{1}$$

Let $\pi_w^\star = \arg\max_\pi J_r(\pi; w)$, $f(w) = J_r(\pi_w^\star; w)$. Our ultimate objective is to find $w^\star = \arg\max_{w \in \mathcal{W}} f(w)$, and to recover a policy $\hat{\pi}_{w^\star}$ such that $J_r(\hat{\pi}_{w^\star}; w^\star)$ is close to $f(w^\star)$. Because $\mathcal{W}$ can be combinatorially large, exhaustive evaluation of all workflows is infeasible. Therefore, we view $f : \mathcal{W} \to \mathbb{R}$ as a black-box function objective that can only be accessed through noisy evaluations obtained by running a workflow-conditioned RL procedure for a fixed budget of environment episodes.

This problem is naturally bi-level: the outer level searches over workflows $w \in \mathcal{W}$ based on past evaluations, while the inner level learns or improves the corresponding workflow-conditioned policy $\pi(\cdot \mid \cdot, w)$ using environment interactions. In Section 4, we introduce the GP-UCB-WS algorithm, which instantiates this bi-level structure using a GP surrogate over workflows and controlled inner-loop policy training.

4. Methodology

We now introduce the workflow search RL formulation in Section 3 with a bi-level algorithm, Gaussian process upper confidence bound workflow search (GP-UCB-WS). The outer loop treats the expected performance of a fixed inner training protocol as a black-box function over the workflow domain and uses an upper confidence bound (UCB) rule to select workflows. The inner loop runs a standard RL algorithm to optimize a workflow-conditioned policy, where we instantiate PPO (Schulman et al., 2017) over a fixed training budget. A key design choice is the use of potential-based reward shaping (PBRS) (Ng et al., 1999; Wiewiora, 2003) during the inner training. The

shaped reward is used only to train the workflow-conditioned policy, while the outer loop is always updated using the original environment reward. PBRS preserves the set of optimal policies, and with an appropriate anchoring of the potential it equalizes shaped and original value functions. As a result, the outer objective over workflows coincides with $f(w)$ in Section 3, even though the inner loop uses a shaped reward to improve adherence to the selected workflow.

4.1. Potential-Based Reward Shaping for Workflow-Conditioned Policies

The embedding ϕ represents the order of the milestones that defines each workflow. We parameterize a workflow-conditioned policy $\pi_\theta(a \mid s, w)$ as a neural network that receives both the state s and a workflow embedding feature $\phi(w)$ as inputs. During training on a given workflow w, we replace original reward r with a shaped reward in (2):

$$r'_w(s, a, s') \;=\; r(s, a, s') \;+\; \lambda\big(\gamma\, \Phi_w(s') - \Phi_w(s)\big), \qquad \lambda \geq 0, \tag{2}$$

where $\Phi_w : \mathcal{S} \to \mathbb{R}$ is a bounded potential capturing progress toward completing workflow w. In practice, the state s is augmented with a progress flag that records which milestones of w have been completed, and Φ_w is defined in terms of this progress signal. We choose weight λ so that shaping signals are informative but do not dominate the environment reward. Under standard PBRS conditions and anchoring, this transformation leaves optimal policies and values unchanged under r. We exploit it only to make the inner-loop optimization better aligned with the prescribed workflow.

4.2. Inner Loop: Fixed-Budget Policy Optimization

Given a selected workflow w_t, the inner loop runs PPO for a fixed number M of updates with K parallel environments, training under r'_w from (2) while always recording original returns. At update $m \in \{1, \ldots, M\}$ we collect K on-policy episodes $\tilde{\tau}^{(i)}_{m,t} := \big(s^{(i)}_{0,m,t}, a^{(i)}_{0,m,t}, s^{(i)}_{1,m,t}, \ldots, a^{(i)}_{T^{(i)}_{m,t}-1,m,t}, s^{(i)}_{T^{(i)}_{m,t},m,t}\big)$ with the current policy $\pi_{m-1}(\cdot \mid \cdot, w_t)$ trained under r'_w, and record the original episodic returns $G_r\big(\tilde{\tau}^{(i)}_{m,t}\big) := \sum_{u=0}^{T^{(i)}_{m,t}-1} \gamma^u\, r\big(s^{(i)}_{u,m,t}, a^{(i)}_{u,m,t}, s^{(i)}_{u+1,m,t}\big)$, where $T^{(i)}_{m,t}$ is the episode length. One PPO update yields π_m. We maintain the per-update mean original return $\hat{J}_r(\pi_m; w_t) = \frac{1}{K}\sum_{i=1}^{K} G_r(\tilde{\tau}^{(i)}_{m,t})$ and, after M updates, select the best checkpoint

$$m^*(w_t) \;\in\; \arg\max_{m \in \{1,\ldots,M\}} \hat{J}_r(\pi_m; w_t), \qquad \pi_{w_t} \leftarrow \pi_{m^*}.$$

The outer-loop observation is the best-iterate mean original return as an empirical estimate of the fixed-budget performance:

$$y_t \;=\; \max_{m \in \{1,\ldots,M\}} \hat{J}_r(\pi_m; w_t) \;=\; \hat{J}_r(\pi_{m^*}; w_t), \tag{3}$$

Throughout, reward shaping is used only for inner-loop training; workflow selection and all reported performance use the original environment reward. Using the best mean original return among the M PPO checkpoints avoids a separate evaluation phase, and the resulting optimism is explicitly captured by the selection-bias term in Lemma 3.

4.3. Outer Loop: Gaussian Process Surrogates and Uncertainty-Guided Workflow Selection

For a fixed inner training budget M, we define the fixed-budget target as

$$g_M(w) \;:=\; \mathbb{E}\Big[\, \max_{1\le m\le M} J_r\big(\pi_m; w\big) \,\Big], \tag{4}$$

the expected original-reward performance of the best checkpoint among the M inner updates on workflow w over both training randomness and environment stochasticity. We treat $g_M : \mathcal{W} \to \mathbb{R}$ as an unknown function on the finite domain and place a GP prior $\mathcal{GP}(0, k_{\mathcal{W}})$ using the workflow kernel from Section 3. At round t, the observation y_t in (3) provides a noisy estimate of $g_M(w_t)$.

After $(t-1)$ observations $\{(w_\tau, y_\tau)\}_{\tau=1}^{t-1}$, we compute the GP posterior $(\mu_{t-1}, \sigma_{t-1})$ on $\mathcal{W}$ and select the next workflow using a finite-domain GP-UCB rule:

$$w_t \;\in\; \arg\max_{w\in\mathcal{W}} \; \mu_{t-1}(w) \;+\; \sqrt{\beta_t}\,\sigma_{t-1}(w), \tag{5}$$

with an exploration schedule β_t chosen following standard prescriptions.

We use a GP surrogate rather than the raw observations y_t because each y_t is only a single noisy fixed-budget estimate of one workflow. Through $k_{\mathcal{W}}$, the GP shares information across structurally similar workflows, which is crucial when $|\mathcal{W}|$ is combinatorial and only a small subset can be evaluated. The posterior variance $\sigma_{t-1}(w)$ then supplies the uncertainty term used by UCB to balance exploration and exploitation. We then run the inner loop on w_t, compute y_t via (3), and update the GP with (w_t, y_t).

4.4. GP-UCB-WS Algorithm

Algorithm 1 summarizes the full procedure. The outer loop maintains a GP surrogate over g_M and selects workflows using GP-UCB; the inner loop applies PPO with PBRS under a fixed interaction budget and returns the scalar y_t to the outer loop.

5. Convergence Analysis

We analyze GP-UCB-WS when the inner learner uses PBRS and the outer observation is the best-iterate mean original return across the M inner updates. The outer objective is $f(w) := \max_\pi J_r(\pi; w)$, and the fixed-budget target is

$$g_M(w) \;:=\; \mathbb{E}\Big[\, \max_{1\le m\le M} J_r(\pi_m; w) \Big].$$

The end-to-end result only requires a uniform inner-loop accuracy bound $0 \le f(w) - g_M(w) \le \Delta_M$ for all $w \in \mathcal{W}$; Lemma 2 instantiates this with $\Delta_M = C_{\mathrm{ppo}}/\sqrt{M}$ under Assumptions A1–A6. We assume $k_{\mathcal{W}}$ is positive semidefinite and rescaled so that $k_{\mathcal{W}}(w, w) \le 1$, and write $K_A^{(\mathcal{W})}$ for the Gram matrix of a queried sequence $A = (w_1, \ldots, w_T)$.

High-level proof sketch. (i) PBRS preserves the outer objective. (ii) The inner learner incurs a uniform gap Δ_M, with PPO giving $\Delta_M = O(M^{-1/2})$. (iii) The best-of-M observation decomposes into sub-Gaussian noise plus bounded optimism bias. (iv) Finite-domain GP-UCB then gives regret to g_M of order $\tilde{O}(\sqrt{T})$ plus $O(TB_{M,K,\delta})$. (v) Adding the inner-loop gap gives Theorem 2.

Algorithm 1: GP-UCB-WS

Input : Workflow set $\mathcal{W}$; kernel $k_{\mathcal{W}}$; shaping weight λ; PPO updates M; parallel envs K
Output : Selected workflow w^* and final policy π_{w^*}
Initialize GP prior $\mathcal{GP}(0, k_{\mathcal{W}})$ over $g_M : \mathcal{W} \to \mathbb{R}$
for $t = 1$ **to** N **do**
  Compute GP posterior $(\mu_{t-1}, \sigma_{t-1})$ on $\mathcal{W}$
  $w_t \leftarrow \arg\max_{w \in \mathcal{W}} \ \mu_{t-1}(w) \ + \ \sqrt{\beta_t}\, \sigma_{t-1}(w)$
  Initialize π_0
  for $m = 1$ **to** M **do**
    for $i = 1$ **to** K **do**
      Sample one on-policy episode with $\pi_{m-1}(\cdot \mid \cdot, w_t)$: $s^{(i)}_{0,m,t} \sim \rho_0$; $a^{(i)}_{u,m,t} \sim \pi_{m-1}(\cdot \mid s^{(i)}_{u,m,t}, w_t)$; $s^{(i)}_{u+1,m,t} \sim P(\cdot \mid s^{(i)}_{u,m,t}, a^{(i)}_{u,m,t})$ for $u = 0, \ldots, T^{(i)}_{m,t} - 1$
      Compute original-return $G_r\big(\tilde{\tau}^{(i)}_{m,t}\big) \leftarrow \sum_{u=0}^{T^{(i)}_{m,t}-1} \gamma^u\, r\big(s^{(i)}_{u,m,t}, a^{(i)}_{u,m,t}, s^{(i)}_{u+1,m,t}\big)$.
    end
    $\hat{J}_r(\pi_m; w_t) \leftarrow \dfrac{1}{K} \sum_{i=1}^{K} G_r\big(\tilde{\tau}^{(i)}_{m,t}\big)$
    $\pi_m \leftarrow \text{PPO-UPDATE}\Big(\pi_{m-1},\ \{\tilde{\tau}^{(i)}_{m,t}\}_{i=1}^{K};\ r'_{w_t}\Big)$
  end
  $m^* \leftarrow \arg\max_{m \in \{1,\ldots,M\}} \hat{J}_r(\pi_m; w_t)$
  $\pi_{w_t} \leftarrow \pi_{m^*}$
  $y_t \leftarrow \hat{J}_r(\pi_{m^*}; w_t)$
  Update GP with datum $(w_t,\ y_t)$
end
return $w^* \leftarrow \arg\max_{w \in \mathcal{W}} \mu_N(w)$ *and* π_{w^*}

Practical scope of the PPO assumptions. Assumptions A1–A6 are sufficient conditions used to instantiate Δ_M, not exact claims about every implementation detail. In our benchmarks, the action spaces are discrete and episodic returns are bounded; the remaining assumptions are standard idealizations, and any mismatch is absorbed into a larger effective Δ_M.

We demonstrate the key lemmas and the main theorems here, and the detailed proofs are in Appendix A.

5.1. Key lemmas

Lemma 1 (PBRS training preserves the outer objective) *For $r'_w(s, a, s') = r(s, a, s') + \lambda(\gamma\Phi_w(s') - \Phi_w(s))$ with bounded Φ_w and $\gamma \in (0, 1)$, and anchoring $\mathbb{E}_{s_0 \sim \rho_0}[\Phi_w(s_0)] = 0$, we have for any policy π: (i) $J_{r'_w}(\pi; w) = J_r(\pi; w)$; (ii) $\arg\max_\pi J_{r'_w}(\pi; w) = \arg\max_\pi J_r(\pi; w)$.*

Lemma 2 (PPO instantiation of the inner fixed-budget accuracy term) *Under Assumptions A1–A6 in Appendix A, the PPO inner loop satisfies*

$$0 \le f(w) - g_M(w) \le \Delta_M, \qquad \Delta_M = \frac{C_{\text{ppo}}}{\sqrt{M}}.$$

Anchored PBRS (Lemma 1) ensures this bound holds for original returns. More generally, Theorem 2 only requires any uniform bound of the form $0 \le f(w) - g_M(w) \le \Delta_M$ on $\mathcal{W}$.

Lemma 3 (Sub-Gaussian noise and selection bias under best-of-M) *Let per-episode returns satisfy $|G_r| \le V$ a.s., and suppose run-to-run training randomness is independent across rounds with*

sub-Gaussian proxy σ_{train}. *At round* t, *with* $y_t = \max_m \hat{J}_r(\pi_m; w_t)$ *and* $g_M(w) = \mathbb{E}[\max_m J_r(\pi_m; w)]$, *we can write* $y_t = g_M(w_t) + \varepsilon_t + b_t$ *where* ε_t *is conditionally mean-zero and* R*-sub-Gaussian with*

$$R \;\le\; \sqrt{\sigma_{\text{train}}^2 \;+\; \frac{V^2}{K}}\;,$$

and, with probability at least $1-\delta$, *the selection-bias term satisfies the uniform bound*

$$0 \;\le\; b_t \;\le\; B_{M,K,\delta} \;:=\; \frac{V}{\sqrt{K}}\sqrt{2\log\Big(\frac{M\,\pi^2 T^2}{6\delta}\Big)} \qquad \textit{for all } t \in \{1,\dots,T\}.$$

Lemma 4 (Simultaneous confidence with bounded bias, finite domain) *Under a GP prior with kernel* k_W *and likelihood variance* R^2, *for* $\beta_t = 2\log\big(|\mathcal{W}|\pi^2 t^2/(6\delta)\big)$,

$$\Pr\Big[\;\forall t \ge 1,\; \forall w \in \mathcal{W}:\; |g_M(w) - \mu_{t-1}(w)| \le \sqrt{\beta_t}\,\sigma_{t-1}(w) \;+\; B_{M,K,\delta}\;\Big] \;\ge\; 1-\delta.$$

Lemma 5 (Instantaneous regret bound with bias) *On the event of Lemma 4, the UCB choice* $w_t \in \arg\max_w \mu_{t-1}(w) + \sqrt{\beta_t}\sigma_{t-1}(w)$ *satisfies*

$$r_t^{(g)} \;:=\; g_M(w_M^\star) - g_M(w_t) \;\le\; 2\sqrt{\beta_t}\,\sigma_{t-1}(w_t) \;+\; 2\,B_{M,K,\delta}, \quad w_M^\star \in \arg\max_w g_M(w).$$

Lemma 6 (Information gain and variance sum) *For the queried sequence* $A = (w_1,\dots,w_T)$,

$$I(y_{1:T}; g_M) \;=\; \tfrac{1}{2}\log\det\big(I + R^{-2}K_A^{(\mathcal{W})}\big) \;=\; \tfrac{1}{2}\sum_{t=1}^{T}\log\big(1 + R^{-2}\sigma_{t-1}^2(w_t)\big),$$

and

$$\sum_{t=1}^{T}\sigma_{t-1}^2(w_t) \;\le\; \frac{2}{\log(1+R^{-2})}\,\gamma_T.$$

5.2. Main theorems

Theorem 1 (Outer regret to the fixed-budget target g_M**)** *With probability at least* $1-\delta$,

$$\sum_{t=1}^{T}\big(g_M(w_M^\star) - g_M(w_t)\big) \;\le\; \sqrt{\frac{8}{\log(1+R^{-2})}\,T\,\beta_T\,\gamma_T} \;+\; 2T\,B_{M,K,\delta}.$$

Sketch. *Lemma 4 gives simultaneous confidence with a bias envelope. The UCB choice yields* $r_t^{(g)} \le 2\sqrt{\beta_t}\,\sigma_{t-1}(w_t) + 2B_{M,K,\delta}$ *(Lemma 5). Sum, apply Cauchy–Schwarz and Lemma 6, with the bias contributing* $2T\,B_{M,K,\delta}$ *gives the result.*

Theorem 2 (End-to-end regret to the original objective f**)** *Under PBRS anchoring and any uniform fixed-budget inner-loop accuracy bound* $0 \le f(w) - g_M(w) \le \Delta_M$ *for all* $w \in \mathcal{W}$, *with probability at least* $1-\delta$,

$$\sum_{t=1}^{T}\big(f(w^\star) - f(w_t)\big) \;\le\; \sqrt{\frac{8}{\log(1+R^{-2})}\,T\,\beta_T\,\gamma_T} \;+\; 2T\,\Delta_M \;+\; 2T\,B_{M,K,\delta}.$$

For the PPO inner loop under Assumptions A1–A6, Lemma 2 gives $\Delta_M = C_{\text{ppo}}/\sqrt{M}$.

Sketch. *For any* t, $f(w^\star) - f(w_t) = \big(g_M(w^\star) - g_M(w_t)\big) + \big(f(w^\star) - g_M(w^\star)\big) + \big(g_M(w_t) - f(w_t)\big) \le r_t^{(g)} + 2\Delta_M$. *Sum and use Theorem 1 lead to the final regret.*

6. Experiments

Grid-world: planning ordered visit with obstacles We evaluate structured workflow search in a 30×30 grid-world with randomly generated obstacles. The task is to visit four 4×4 corner regions in a specified order requiring crossing the main diagonal twice. The action space is discrete movements towards four directions, and the observations include the agent's coordinates (x, y). Each episode ends after 500 steps or once all checkpoints have been visited in the required order. The reward enforces strict adherence to the current workflow, with a small per-step penalty to encourage short paths; thus, the optimal return is achieved only by visiting the target sequence while avoiding blocked cells.

TTCP CAGE Challenge 2 We evaluate workflow search on the TTCP CAGE Challenge 2 cyber defense task (Kiely et al., 2023). The enterprise network is divided into user, enterprise, and operational subnets, each containing hosts and servers. A workflow is a priority ordering over five asset classes, specifying which assets to fix first during an attack. The defender's observation includes local alerts, host state and connectivity, with the workflow embedding appended to the state. We evaluate against two built-in red agents: B-line, which drives directly toward the Operational Server using layout knowledge, and Meander, which escalates privilege subnet-by-subnet before targeting the Operational Server. Red agents select exploits from a ranked list conditioned on discovered open ports. The reward is composed of penalties for service disruptions to normal user activity and for successful attacker compromises.

Learning protocol During training, PBRS rewards progress along the sampled workflow and penalizes regressions. The policy is conditioned on the workflow and trained with PPO under a fixed interaction budget. The outer-loop observation is the highest mean original return among the M PPO checkpoints, matching (3). Reward shaping is used only for inner-loop training; workflow selection and all reported results use the original environment reward.

Table 1 summarizes the training performance with averaged episodic reward levels across three runs on the gridworld and the CAGE Challenge 2 against both B_line attacker and Meander attacker, while Figure 1 shows the learning curves for each scenario. For the ordered visit planning task in the gridworld, only GP-UCB-WS identifies the correct visit order and achieves a positive return, while PPO, HRL, and SAC fail due to the exploration complexity without a structured representation of candidate plans. For CAGE Challenge 2, GP-UCB-WS achieves faster early gains and comparable final returns to the baselines. Against B-line, GP-UCB-WS closes the initial performance gap within the first few thousand episodes and continues to improve, reflecting the benefit of searching for robust priority orders under a moving threat. Against Meander, the advantage persists throughout training, suggesting that adaptive workflow selection helps the defender prioritize and re-prioritize subsystems until the optimal mode is discovered.

7. Conclusion

We proposed workflow search RL and the GP-UCB-WS algorithm. The method couples a GP-UCB outer loop over workflow embeddings with an inner workflow-conditioned policy, trained using anchored potential-based reward shaping and evaluated on the original reward. This separates combinatorial ordering from continuous control, accelerates learning, and yields interpretable outer-loop traces. On an ordered-visit gridworld and the TTCP CAGE Challenge 2, the approach outperforms

Table 1: Episodic reward after the indicated number of training episodes for Gridworld ordered-visit with obstacles planning task and CAGE Challenge 2 defender performance vs. B-line and Meander attackers. Results are averaged over three runs. Our GP-UCB-WS outperforms all the baselines across three different scenarios with higher rewards at different stages of training.

Episodes	GP-UCB-WS (Ours)	PPO	HRL	SAC
Gridworld Planning				
10000	-4.24 ± 1.07	$\mathbf{-1.77 \pm 1.64}$	-4.71 ± 0.13	-4.92 ± 0.01
20000	$\mathbf{0.74 \pm 4.09}$	-1.13 ± 1.48	-4.83 ± 0.04	-4.95 ± 0.02
30000	$\mathbf{4.01 \pm 0.42}$	-0.47 ± 0.29	-4.75 ± 0.10	-5.00 ± 0.00
40000	$\mathbf{4.27 \pm 0.23}$	-0.71 ± 0.43	-4.83 ± 0.15	-5.00 ± 0.00
50000	$\mathbf{4.11 \pm 0.37}$	-0.21 ± 0.36	-4.73 ± 0.10	-5.00 ± 0.00
CAGE2 — B-line				
5000	$\mathbf{-30.46 \pm 2.82}$	-180.53 ± 4.80	-138.02 ± 2.41	-60.59 ± 1.42
7500	$\mathbf{-26.69 \pm 3.10}$	-122.02 ± 3.99	-95.36 ± 10.65	-61.87 ± 2.35
10000	$\mathbf{-29.10 \pm 3.71}$	-78.20 ± 13.17	-58.18 ± 8.43	-53.96 ± 6.01
12500	$\mathbf{-29.58 \pm 6.07}$	-34.48 ± 2.54	-47.21 ± 4.04	-53.39 ± 1.72
15000	$\mathbf{-23.91 \pm 3.47}$	-30.03 ± 1.86	-30.60 ± 3.36	-51.47 ± 3.79
17500	$\mathbf{-21.52 \pm 4.16}$	-27.29 ± 1.73	-27.96 ± 0.60	-47.14 ± 1.48
20000	$\mathbf{-23.07 \pm 2.30}$	-24.33 ± 0.71	-26.59 ± 1.35	-46.03 ± 1.40
CAGE2 — Meander				
5000	$\mathbf{-36.14 \pm 1.59}$	-148.26 ± 15.26	-148.52 ± 10.90	-50.51 ± 2.68
7500	$\mathbf{-32.06 \pm 1.91}$	-104.28 ± 6.27	-85.88 ± 21.04	-53.80 ± 3.13
10000	$\mathbf{-29.27 \pm 1.42}$	-64.29 ± 8.53	-55.43 ± 3.93	-50.43 ± 2.41
12500	$\mathbf{-26.89 \pm 1.55}$	-44.56 ± 7.37	-44.66 ± 4.07	-52.44 ± 2.85
15000	$\mathbf{-27.50 \pm 1.38}$	-36.30 ± 4.21	-38.03 ± 2.39	-51.60 ± 2.46
17500	$\mathbf{-27.30 \pm 2.04}$	-31.31 ± 2.12	-32.33 ± 1.32	-50.87 ± 0.54
20000	$\mathbf{-25.14 \pm 0.89}$	-26.84 ± 2.27	-28.40 ± 0.92	-49.24 ± 1.70

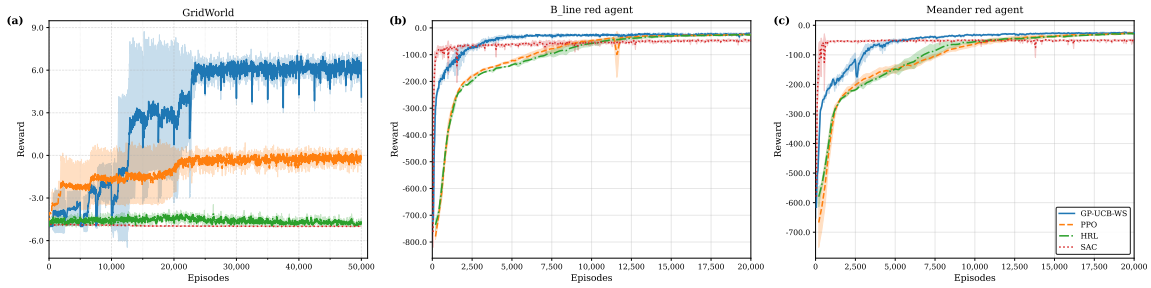

Figure 1: This figure shows the reward change with training episodes for both the Grid-world environment and CAGE Challenge 2 for Bline and Meander red agents. Mean returns are shown across three runs. GP-UCB-WS is the only method that discovers the desired order of visits in the gridworld, outperforming PPO, HRL, and SAC. For CAGE Challenge 2, GP-UCB-WS exhibits faster early learning against both attackers and maintains a consistent advantage against other methods.

flat and hierarchical baselines, highlighting the value of adaptive outer-loop search over structured decompositions.

Acknowledgments

This work was supported in part by the United States Military Academy (USMA) under Cooperative Agreement No. W911NF-23-2-0175. The views and conclusions expressed in this paper are those of the authors and do not reflect the official policy or position of the U.S. Military Academy, U.S. Army, U.S. Department of Defense, or U.S. Government.

References

Jacob Andreas, Dan Klein, and Sergey Levine. Modular multitask reinforcement learning with policy sketches. In *International conference on machine learning*, pages 166–175. PMLR, 2017.

Pierre-Luc Bacon, Jean Harb, and Doina Precup. The option-critic architecture. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

Aryan Deshwal and Jana Doppa. Combining latent space and structured kernels for bayesian optimization over combinatorial spaces. *Advances in neural information processing systems*, 34: 8185–8200, 2021.

Thomas G Dietterich. Hierarchical reinforcement learning with the maxq value function decomposition. *Journal of artificial intelligence research*, 13:227–303, 2000.

Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. In *International Conference on Learning Representations*.

Gregory Farquhar, Tim Rocktäschel, Maximilian Igl, and Shimon Whiteson. Treeqn and atreec: Differentiable tree-structured models for deep reinforcement learning. *arXiv preprint arXiv:1710.11417*, 2017.

Karol Gregor, Danilo Jimenez Rezende, and Daan Wierstra. Variational intrinsic control. *arXiv preprint arXiv:1611.07507*, 2016.

Steven Hansen, Will Dabney, Andre Barreto, David Warde-Farley, Tom Van de Wiele, and Volodymyr Mnih. Fast task inference with variational intrinsic successor features. In *International Conference on Learning Representations*.

Nai-Chieh Huang, Ping-Chun Hsieh, Kuo-Hao Ho, and I-Chen Wu. Ppo-clip attains global optimality: Towards deeper understandings of clipping. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 12600–12607, 2024.

Rodrigo Toro Icarte, Toryn Q Klassen, Richard Valenzano, and Sheila A McIlraith. Reward machines: Exploiting reward function structure in reinforcement learning. *Journal of Artificial Intelligence Research*, 73:173–208, 2022.

Mitchell Kiely, David Bowman, Maxwell Standen, and Christopher Moir. On autonomous agents in a cyber defence environment. *arXiv preprint arXiv:2309.07388*, 2023.

Michael Laskin, Hao Liu, Xue Bin Peng, Denis Yarats, Aravind Rajeswaran, and Pieter Abbeel. Unsupervised reinforcement learning with contrastive intrinsic control. *Advances in Neural Information Processing Systems*, 35:34478–34491, 2022.

Andrew Levy, George Konidaris, Robert Platt, and Kate Saenko. Learning multi-level hierarchies with hindsight. *arXiv preprint arXiv:1712.00948*, 2017.

Ofir Nachum, Shixiang Shane Gu, Honglak Lee, and Sergey Levine. Data-efficient hierarchical reinforcement learning. *Advances in neural information processing systems*, 31, 2018.

Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, volume 99, pages 278–287. Citeseer, 1999.

Changyong Oh, Jakub Tomczak, Efstratios Gavves, and Max Welling. Combinatorial bayesian optimization using the graph cartesian product. *Advances in Neural Information Processing Systems*, 32, 2019.

Seohong Park, Jongwook Choi, Jaekyeom Kim, Honglak Lee, and Gunhee Kim. Lipschitz-constrained unsupervised skill discovery. In *International Conference on Learning Representations*.

Seohong Park, Kimin Lee, Youngwoon Lee, and Pieter Abbeel. Controllability-aware unsupervised skill discovery. *arXiv preprint arXiv:2302.05103*, 2023.

Sébastien Racanière, Théophane Weber, David Reichert, Lars Buesing, Arthur Guez, Danilo Jimenez Rezende, Adrià Puigdomènech Badia, Oriol Vinyals, Nicolas Heess, Yujia Li, et al. Imagination-augmented agents for deep reinforcement learning. *Advances in neural information processing systems*, 30, 2017.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Archit Sharma, Shixiang Gu, Sergey Levine, Vikash Kumar, and Karol Hausman. Dynamics-aware unsupervised discovery of skills. In *International Conference on Learning Representations*.

Richard S Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.

Aviv Tamar, Yi Wu, Garrett Thomas, Sergey Levine, and Pieter Abbeel. Value iteration networks. *Advances in neural information processing systems*, 29, 2016.

Rodrigo Toro Icarte, Ethan Waldie, Toryn Klassen, Rick Valenzano, Margarita Castro, and Sheila McIlraith. Learning reward machines for partially observable reinforcement learning. *Advances in neural information processing systems*, 32, 2019.

Abhinav Verma, Vijayaraghavan Murali, Rishabh Singh, Pushmeet Kohli, and Swarat Chaudhuri. Programmatically interpretable reinforcement learning. In *International conference on machine learning*, pages 5045–5054. PMLR, 2018.

Alexander Sasha Vezhnevets, Simon Osindero, Tom Schaul, Nicolas Heess, Max Jaderberg, David Silver, and Koray Kavukcuoglu. Feudal networks for hierarchical reinforcement learning. In *International conference on machine learning*, pages 3540–3549. PMLR, 2017.

Eric Wiewiora. Potential-based shaping and q-value initialization are equivalent. *Journal of Artificial Intelligence Research*, 19:205–208, 2003.

Yang Yu. Reinforcement learning with anticipation: A hierarchical approach for long-horizon tasks. *arXiv preprint arXiv:2509.05545*, 2025.

Yuan Zhuang, Yi Shen, Zhili Zhang, Yuxiao Chen, and Fei Miao. Yolo-marl: You only llm once for multi-agent reinforcement learning. In *ICRA 2025 Workshop on Foundation Models and Neuro-Symbolic AI for Robotics*.

Appendix A. Proofs and Additional Details for Section 5

Lemma 1 (PBRS training preserves the outer objective) *For $r'_w(s,a,s') = r(s,a,s') + \lambda(\gamma\Phi_w(s') - \Phi_w(s))$ with bounded Φ_w and $\gamma \in (0,1)$, and anchoring $\mathbb{E}_{s_0\sim\rho_0}[\Phi_w(s_0)] = 0$, we have for any policy π: (i) $J_{r'_w}(\pi; w) = J_r(\pi; w)$; (ii) $\arg\max_\pi J_{r'_w}(\pi; w) = \arg\max_\pi J_r(\pi; w)$.*

Proof Consider any trajectory $(s_0, a_0, s_1, a_1, \dots)$ and the shaped reward

$$r'_w(s,a,s') \;=\; r(s,a,s') \;+\; \lambda\big(\gamma\,\Phi_w(s') - \Phi_w(s)\big). \tag{6}$$

Telescoping identity. Define the partial sums

$$S_T \;:=\; \sum_{t=0}^{T} \gamma^t\big(\gamma\Phi_w(s_{t+1}) - \Phi_w(s_t)\big). \tag{7}$$

Split the two series and re-index the first:

$$S_T \;=\; \sum_{t=0}^{T} \gamma^{t+1}\Phi_w(s_{t+1}) \;-\; \sum_{t=0}^{T} \gamma^t\Phi_w(s_t). \tag{8}$$

Extract the $t = 0$ term from the second sum and cancel the overlapping terms:

$$S_T \;=\; \Big(\sum_{t=1}^{T+1} \gamma^t\Phi_w(s_t)\Big) \;-\; \Big(\Phi_w(s_0) + \sum_{t=1}^{T} \gamma^t\Phi_w(s_t)\Big) \;=\; -\,\Phi_w(s_0) \;+\; \gamma^{T+1}\Phi_w(s_{T+1}). \tag{9}$$

If Φ_w is bounded, say $|\Phi_w| \le B < \infty$, then the tail vanishes as $T \to \infty$ because $\gamma \in (0,1)$:

$$\lim_{T\to\infty} \gamma^{T+1}\Phi_w(s_{T+1}) \;=\; 0. \tag{10}$$

Thus the infinite series telescopes to

$$\sum_{t=0}^{\infty} \gamma^t\big(\gamma\Phi_w(s_{t+1}) - \Phi_w(s_t)\big) \;=\; \lim_{T\to\infty} S_T \;=\; -\,\Phi_w(s_0). \tag{11}$$

Equality of returns under anchoring. Sum the shaped reward along the trajectory and separate the shaping term:

$$\sum_{t=0}^{\infty} \gamma^t\, r'_w(s_t,a_t,s_{t+1}) \;=\; \sum_{t=0}^{\infty} \gamma^t\, r(s_t,a_t,s_{t+1}) \;+\; \lambda \sum_{t=0}^{\infty} \gamma^t\big(\gamma\Phi_w(s_{t+1}) - \Phi_w(s_t)\big). \tag{12}$$

Apply (11):

$$\sum_{t=0}^{\infty} \gamma^t\, r'_w(s_t,a_t,s_{t+1}) \;=\; \sum_{t=0}^{\infty} \gamma^t\, r(s_t,a_t,s_{t+1}) \;-\; \lambda\,\Phi_w(s_0). \tag{13}$$

Taking expectations over trajectories generated by π and P, with $s_0 \sim \rho_0$,

$$J_{r'_w}(\pi; w) \;=\; J_r(\pi; w) \;-\; \lambda\,\mathbb{E}_{s_0\sim\rho_0}\big[\Phi_w(s_0)\big]. \tag{14}$$

Under the anchoring condition $\mathbb{E}_{s_0\sim\rho_0}[\Phi_w(s_0)] = 0$, we obtain

$$J_{r'_w}(\pi; w) \;=\; J_r(\pi; w) \quad \text{for all } \pi. \tag{15}$$

This proves the preserved outer objective (part (i)).

Value/advantage invariance (part (ii)). Write the Bellman equation for Q'_π:

$$Q'_\pi(s,a) \;=\; \mathbb{E}\big[\, r'_w(s,a,s') \;+\; \gamma\, V'_\pi(s') \,\big]. \tag{16}$$

Expand r'_w and regroup terms:

$$\begin{aligned} Q'_\pi(s,a) &= \mathbb{E}\big[\, r(s,a,s') \;+\; \lambda(\gamma\Phi_w(s') - \Phi_w(s)) \;+\; \gamma\, V'_\pi(s') \,\big] && (17)\\ &= \underbrace{\mathbb{E}\big[\, r(s,a,s') + \gamma V_\pi(s') \,\big]}_{=\, Q_\pi(s,a)} \;+\; \lambda\Big(\gamma\, \mathbb{E}[\Phi_w(s')] - \Phi_w(s)\Big) \;+\; \gamma\, \mathbb{E}\big[\, V'_\pi(s') - V_\pi(s') \,\big]. && (18) \end{aligned}$$

Suppose V'_π satisfies

$$V'_\pi(s) \;=\; V_\pi(s) \;-\; \lambda\, \Phi_w(s). \tag{19}$$

Then

$$\mathbb{E}\big[V'_\pi(s') - V_\pi(s')\big] \;=\; -\,\lambda\, \mathbb{E}[\Phi_w(s')]. \tag{20}$$

Substitute into (18):

$$\begin{aligned} Q'_\pi(s,a) &= Q_\pi(s,a) \;+\; \lambda\Big(\gamma\, \mathbb{E}[\Phi_w(s')] - \Phi_w(s)\Big) \;-\; \gamma\, \lambda\, \mathbb{E}[\Phi_w(s')] && (21)\\ &= Q_\pi(s,a) \;-\; \lambda\, \Phi_w(s). && (22) \end{aligned}$$

Taking expectation over $a \sim \pi(\cdot \mid s, w)$ gives

$$V'_\pi(s) \;=\; \mathbb{E}_{a\sim\pi(\cdot|s,w)}[\, Q'_\pi(s,a) \,] \;=\; \mathbb{E}_{a\sim\pi(\cdot|s,w)}[\, Q_\pi(s,a) - \lambda\Phi_w(s) \,] \;=\; V_\pi(s) \;-\; \lambda\, \Phi_w(s), \tag{23}$$

which verifies (19). By uniqueness of the Bellman fixed point, (22)–(23) characterize Q'_π, V'_π. Finally, the advantages coincide:

$$A'_\pi(s,a) \;=\; Q'_\pi(s,a) - V'_\pi(s) \;=\; \big(Q_\pi(s,a) - \lambda\Phi_w(s)\big) - \big(V_\pi(s) - \lambda\Phi_w(s)\big) \;=\; A_\pi(s,a). \tag{24}$$

Hence the greedy action sets are identical ($\arg\max_a Q'_\pi(s,a) \;=\; \arg\max_a Q_\pi(s,a)$), which implies policy invariance. ■

Lemma 2 (PPO instantiation of the inner fixed-budget accuracy term) *Under the standard assumptions used in the neural PPO-Clip analysis of Huang et al. (2024), the PPO inner loop satisfies*

$$0 \le f(w) - g_M(w) \le \Delta_M, \qquad \Delta_M = \frac{C_{\mathrm{ppo}}}{\sqrt{M}}.$$

Anchored PBRS (Lemma 1) ensures this bound holds for original returns. The assumptions below should be interpreted as sufficient conditions for this PPO instantiation of the generic inner-loop term Δ_M; the end-to-end result in Theorem 2 only requires a uniform bound $0 \le f(w) - g_M(w) \le \Delta_M$.

Assume the following conditions from Huang et al. (2024) hold (notation as in their paper):

(A1) Finite actions and bounded rewards. *$|\mathcal{A}| < \infty$, $\gamma \in (0,1)$, and $|r(s,a)| \le R_{\max}$.*

(A2) Function-class realizability (Assumption 1). *For any policy π, its Q^π lies in a sufficiently wide two-layer neural class and the class is closed under the Bellman operator T_π.*

(A3) Regularity of discounted visitation (Assumption 4). *A small-ball / density regularity condition holds for the policy-induced discounted state–action measure, yielding sharp regression/concentration bounds.*

(A4) Concentrability (Assumption 5). *The concentrability coefficients φ^*, ψ^*, and the global constant C_∞ are finite.*

(A5) PPO-Clip classifier bounds (Equations (14) and (15)). *Let $C_t(s,a)$ denote the generalized PPO-Clip classifier. There exist functions $L_C(T), U_C(T)$ such that for all t,*

$$L_C(T)\,|A^*(s,a)| \;\le\; C_t(s,a)\,|A^*(s,a)| \;\le\; U_C(T)\,|A^*(s,a)|, \qquad L_C(T)=\Omega(T^{-1}),\; U_C(T)=O(T^{-1/2}).$$

(A6) Step sizes, temperatures, and capacity. *Identify $T \equiv M$ and choose entropic mirror descent (EMDA) step size and temperature*

$$\eta \;=\; M^{-1/2}, \qquad \tau_t \;=\; \frac{M^{1/2}}{K\,t},$$

and take the widths (m_f, m_Q) and the number of per-iteration TD/SGD updates T_{upd} large enough that the cumulative regression/evaluation errors satisfy

$$\sum_{t=0}^{M-1} \big(\varepsilon_t + \varepsilon_t'\big) \;=\; O(1).$$

Then there exists a constant $C_{\mathrm{ppo}} > 0$ such that

$$0 \;\le\; f(w) - g_M(w) \;\le\; \frac{C_{\mathrm{ppo}}}{\sqrt{M}}, \qquad g_M(w) \;:=\; \mathbb{E}\Big[\max_{1\le m\le M} J_r(\pi_m; w)\Big].$$

Lemma 3 (Sub-Gaussian noise and selection bias under best-of-M) *Let per-episode returns satisfy $|G_r| \le V$ a.s., and suppose run-to-run training randomness is independent across rounds with sub-Gaussian proxy σ_{train}. At round t, with $y_t = \max_m \hat{J}_r(\pi_m; w_t)$ and $g_M(w) = \mathbb{E}[\max_m J_r(\pi_m; w)]$, we can write $y_t = g_M(w_t) + \varepsilon_t + b_t$ where ε_t is conditionally mean-zero and R-sub-Gaussian with*

$$R \;\le\; \sqrt{\sigma_{\mathrm{train}}^2 \;+\; \frac{V^2}{K}}\;,$$

and, with probability at least $1-\delta$, the selection-bias term satisfies the uniform bound

$$0 \;\le\; b_t \;\le\; B_{M,K,\delta} \;:=\; \frac{V}{\sqrt{K}}\sqrt{2\log\Big(\frac{M\,\pi^2 T^2}{6\delta}\Big)} \qquad \textit{for all } t \in \{1,\dots,T\}.$$

Proof Fix round t and let $\mathcal{F}_{t-1}$ be the sigma-field generated by all data up to round $t-1$. For each inner update $m \in \{1,\dots,M\}$, define the original reward mean return and its empirical error

$$\mu_{m,t} \;:=\; J_r(\pi_m; w_t), \qquad \xi_{m,t} \;:=\; \frac{1}{K}\sum_{i=1}^{K}\Big(G_r\big(\tilde{\tau}_{m,t}^{(i)}\big) \;-\; \mu_{m,t}\Big). \tag{25}$$

By construction,

$$\hat{J}_r(\pi_m; w_t) \;=\; \mu_{m,t} \;+\; \xi_{m,t}. \tag{26}$$

Sub-Gaussianity of each $\xi_{m,t}$. Assume per-episode returns are almost surely bounded: $|G_r(\tilde{\tau}_{m,t}^{(i)})| \le V$. Conditioning on $\mathcal{F}_{t-1}$ and on the random policy π_m, the K summands in (25) are i.i.d. with mean $\mu_{m,t}$ and range in $[-V, V]$; hence, by Hoeffding's lemma, for any $\lambda \in \mathbb{R}$,

$$\mathbb{E}\Big[\exp\big(\lambda\,\xi_{m,t}\big)\;\Big|\;\mathcal{F}_{t-1},\,\pi_m\Big] \;\le\; \exp\!\left(\frac{\lambda^2 V^2}{2K}\right). \tag{27}$$

Thus $\xi_{m,t}$ is conditionally mean-zero and $(V/\sqrt{K})$-sub-Gaussian.

Best-of-M statistic and observation. The outer-loop observation is the maximum over the M checkpoints:

$$y_t \;=\; \max_{1\le m\le M}\; \hat{J}_r(\pi_m; w_t) \;=\; \max_{1\le m\le M}\; \big(\mu_{m,t} + \xi_{m,t}\big). \tag{28}$$

Bias term b_t (nonnegativity). Define the conditional optimism bias

$$b_t \;:=\; \mathbb{E}\big[y_t \;\big|\; \mathcal{F}_{t-1}\big] \;-\; \mathbb{E}\left[\max_{1\le m\le M}\; \mu_{m,t} \;\middle|\; \mathcal{F}_{t-1}\right]. \tag{29}$$

Since the map $x \mapsto \max_m x_m$ is convex and $\mathbb{E}[\xi_{m,t} \mid \mathcal{F}_{t-1}] = 0$, Jensen's inequality yields

$$\mathbb{E}\big[y_t \;\big|\; \mathcal{F}_{t-1}\big] = \mathbb{E}\Big[\max_m\,\big(\mu_{m,t} + \xi_{m,t}\big)\;\Big|\;\mathcal{F}_{t-1}\Big] \;\ge\; \max_m\; \mathbb{E}\Big[\mu_{m,t} + \xi_{m,t}\;\Big|\;\mathcal{F}_{t-1}\Big] \tag{30}$$

$$=\; \max_m\; \mathbb{E}\big[\mu_{m,t} \;\big|\; \mathcal{F}_{t-1}\big]\,, \tag{31}$$

so $b_t \ge 0$.

Centered noise ε_t and sub-Gaussian proxy. Define the centered fluctuation

$$\varepsilon_t \;:=\; y_t \;-\; \mathbb{E}\big[y_t \;\big|\; \mathcal{F}_{t-1}\big]\,. \tag{32}$$

We decompose ε_t into an evaluation-noise component driven by the $\{\xi_{m,t}\}$ and a training-variability component driven by the randomness of $\{\mu_{m,t}\}$ across runs.

(a) Evaluation-noise part. Condition on the (random) vector $\mu_t := (\mu_{1,t}, \ldots, \mu_{M,t})$. Define the function $f_{\mu_t} : \mathbb{R}^M \to \mathbb{R}$ by

$$f_{\mu_t}(x_1, \ldots, x_M) \;:=\; \max_{1\le m\le M}\,\big(\mu_{m,t} + x_m\big). \tag{33}$$

Then $y_t = f_{\mu_t}(\xi_{1,t}, \ldots, \xi_{M,t})$ and f_{μ_t} is 1-Lipschitz with respect to the ℓ_∞-norm:

$$\big|f_{\mu_t}(x) - f_{\mu_t}(y)\big| \;\le\; \|x - y\|_\infty \qquad \text{for all } x, y \in \mathbb{R}^M. \tag{34}$$

Using (27) and a standard concentration result for Lipschitz functionals of independent sub-Gaussians, we obtain, for any $\lambda \in \mathbb{R}$,

$$\mathbb{E}\Big[\exp\Big(\lambda\big(f_{\mu_t}(\xi_{1,t}, \ldots, \xi_{M,t}) - \mathbb{E}[f_{\mu_t}(\xi_{1,t}, \ldots, \xi_{M,t}) \mid \mu_t]\big)\Big)\;\Big|\;\mu_t\Big] \;\le\; \exp\!\left(\frac{\lambda^2 V^2}{2K}\right). \tag{35}$$

That is, the centered evaluation-noise term is $(V/\sqrt{K})$-sub-Gaussian conditionally on μ_t, and hence on $\mathcal{F}_{t-1}$.

(b) Training-variability part. Let

$$Z_t \;:=\; \max_{1\le m\le M} \mu_{m,t} \;-\; \mathbb{E}\Big[\max_{1\le m\le M} \mu_{m,t} \;\Big|\; \mathcal{F}_{t-1}\Big]. \tag{36}$$

By assumption on run-to-run variability, Z_t is conditionally mean-zero and σ_{train}-sub-Gaussian, and is independent of the evaluation noise $\{\xi_{m,t}\}$.

(c) Combine the two parts. Write

$$\begin{aligned}\varepsilon_t &= \; y_t - \mathbb{E}[y_t \mid \mathcal{F}_{t-1}] \\ &= \; \Big(y_t - \mathbb{E}[y_t \mid \mu_t, \mathcal{F}_{t-1}]\Big) \;+\; \Big(\mathbb{E}[y_t \mid \mu_t, \mathcal{F}_{t-1}] - \mathbb{E}[y_t \mid \mathcal{F}_{t-1}]\Big).\end{aligned} \tag{37}$$

The first bracket in (37) is the centered evaluation-noise term from (35), and the second is a centered function of μ_t only, proportional to the centered max in (36). Using independence of the two parts and (35), we obtain the conditional mgf bound

$$\mathbb{E}\Big[\exp\big(\lambda\,\varepsilon_t\big)\;\Big|\;\mathcal{F}_{t-1}\Big] \;\le\; \exp\!\left(\frac{\lambda^2}{2}\Big(\frac{V^2}{K} \;+\; \sigma_{\text{train}}^2\Big)\right), \qquad \lambda\in\mathbb{R}, \tag{38}$$

i.e., ε_t is conditionally mean-zero and R-sub-Gaussian with

$$R \;\le\; \sqrt{\frac{V^2}{K} \;+\; \sigma_{\text{train}}^2}\;. \tag{39}$$

Uniform high-probability envelope for the bias b_t. Let $\alpha > 0$ and consider the event

$$\mathcal{E}_\alpha \;:=\; \bigcap_{t=1}^{T}\bigcap_{m=1}^{M} \big\{\,|\xi_{m,t}| \;\le\; \alpha\,\big\}. \tag{40}$$

By (27) and a union bound over m and t,

$$\Pr\big(\mathcal{E}_\alpha^c\big) \;\le\; 2\,M\,T\;\exp\!\left(-\frac{K\,\alpha^2}{2V^2}\right). \tag{41}$$

Choosing

$$\alpha \;=\; \frac{V}{\sqrt{K}}\,\sqrt{2\log\Big(\frac{M\,\pi^2\,T^2}{6\delta}\Big)}, \tag{42}$$

gives $\Pr(\mathcal{E}_\alpha) \ge 1-\delta$ since $\sum_{t\ge1} 1/t^2 = \pi^2/6$. On $\mathcal{E}_\alpha$, for every t,

$$\max_m \mu_{m,t} \;-\; \alpha \;\le\; \max_m \big(\mu_{m,t} + \xi_{m,t}\big) \;\le\; \max_m \mu_{m,t} \;+\; \alpha, \tag{43}$$

hence by taking conditional expectations given $\mathcal{F}_{t-1}$,

$$0 \;\le\; b_t \;=\; \mathbb{E}[y_t \mid \mathcal{F}_{t-1}] \;-\; \mathbb{E}[\max_m \mu_{m,t} \mid \mathcal{F}_{t-1}] \;\le\; \alpha. \tag{44}$$

Combining (39) and (44) completes the proof. ∎

Lemma 4 (Simultaneous confidence with bounded bias, finite domain) *Under a GP prior with kernel k_W and likelihood variance R^2, for*

$$\beta_t = 2\log\big(|\mathcal{W}|\pi^2 t^2/(6\delta)\big) \tag{45}$$

,

$$\Pr\Big[\,\forall t \geq 1,\ \forall w \in \mathcal{W}:\ |g_M(w) - \mu_{t-1}(w)| \leq \sqrt{\beta_t}\,\sigma_{t-1}(w)\ +\ B_{M,K,\delta}\,\Big]\ \geq\ 1-\delta. \tag{46}$$

Proof Fix $t \geq 1$ and let $\mathcal{F}_{t-1}$ be the sigma-field generated by $\{(w_\tau, y_\tau)\}_{\tau=1}^{t-1}$. Recall the observation decomposition from Lemma 3:

$$y_\tau\ =\ g_M(w_\tau)\ +\ \varepsilon_\tau\ +\ b_\tau, \qquad \text{with } \varepsilon_\tau \text{ conditionally mean-zero and } R\text{-sub-Gaussian}, \tag{47}$$

and the uniform high-probability envelope on the bias

$$\Pr\Big(\max_{1\leq\tau\leq t-1}|b_\tau|\ \leq\ B_{M,K,\delta/2}\Big)\ \geq\ 1-\delta/2. \tag{48}$$

Step 1: De-bias the targets. Define the de-biased pseudo-observations

$$\tilde{y}_\tau\ :=\ y_\tau - b_\tau \quad\Longrightarrow\quad \tilde{y}_\tau\ =\ g_M(w_\tau) + \varepsilon_\tau. \tag{49}$$

Let $\tilde{\mu}_{t-1}$ and $\tilde{\sigma}_{t-1}$ be the GP posterior mean and standard deviation computed from the dataset $\{(w_\tau, \tilde{y}_\tau)\}_{\tau=1}^{t-1}$ under the prior $\mathcal{GP}(0, k_W)$ and Gaussian likelihood with variance R^2. By the *standard finite-domain GP confidence bound* (Gaussian tail + union bound over t, w), with probability at least $1-\delta/2$,

$$\forall t \geq 1,\ \forall w \in \mathcal{W}: \qquad \big|\,g_M(w) - \tilde{\mu}_{t-1}(w)\,\big|\ \leq\ \sqrt{\beta_t}\,\tilde{\sigma}_{t-1}(w), \tag{50}$$

with β_t as in (4).

Step 2: Relate posteriors with and without de-biasing. Write the usual GP/KRR closed form with $A_{t-1} := K_{t-1}^{(W)} + R^2 I$ and $k_{t-1}(w) := [k_W(w, w_1), \ldots, k_W(w, w_{t-1})]^\top$:

$$\tilde{\mu}_{t-1}(w)\ =\ k_{t-1}(w)^\top A_{t-1}^{-1}\,\tilde{y}_{1:t-1}, \qquad \mu_{t-1}(w)\ =\ k_{t-1}(w)^\top A_{t-1}^{-1}\,y_{1:t-1}. \tag{51}$$

Subtract (51) and use $\tilde{y} = y - b$:

$$\mu_{t-1}(w) - \tilde{\mu}_{t-1}(w)\ =\ k_{t-1}(w)^\top A_{t-1}^{-1}\,b_{1:t-1}. \tag{52}$$

Taking absolute values and the ℓ_∞ bound $\|b_{1:t-1}\|_\infty \leq B_{M,K,\delta/2}$ from (48) gives

$$\big|\mu_{t-1}(w) - \tilde{\mu}_{t-1}(w)\big|\ \leq\ \big\|k_{t-1}(w)^\top A_{t-1}^{-1}\big\|_1\,\|b_{1:t-1}\|_\infty\ \leq\ \big\|k_{t-1}(w)^\top A_{t-1}^{-1}\big\|_1\,B_{M,K,\delta/2}. \tag{53}$$

For kernel matrices rescaled so that $k_W(w,w) \leq 1$ and $R^2 > 0$, the GP/KRR smoother is 1-Lipschitz with respect to ℓ_∞ perturbations of the targets, so $\|k_{t-1}(w)^\top A_{t-1}^{-1}\|_1 \leq 1$, and hence

$$\big|\mu_{t-1}(w) - \tilde{\mu}_{t-1}(w)\big|\ \leq\ B_{M,K,\delta/2}. \tag{54}$$

Moreover, the GP posterior variance depends only on inputs (not on targets), so

$$\sigma_{t-1}(w)\ =\ \tilde{\sigma}_{t-1}(w) \qquad \text{for all } w \in \mathcal{W}. \tag{55}$$

Step 3: Triangle inequality and union of events. Combining (50), (54), and (55), we obtain, on the intersection of the two events of probability at least $1 - \delta/2$ each,

$$\begin{aligned} \left| g_M(w) - \mu_{t-1}(w) \right| &\le \left| g_M(w) - \tilde{\mu}_{t-1}(w) \right| \; + \; \left| \tilde{\mu}_{t-1}(w) - \mu_{t-1}(w) \right| && (56) \\ &\le \sqrt{\beta_t}\, \tilde{\sigma}_{t-1}(w) \; + \; B_{M,K,\delta/2} && (57) \\ &= \sqrt{\beta_t}\, \sigma_{t-1}(w) \; + \; B_{M,K,\delta/2}. && (58) \end{aligned}$$

Applying a union bound over the two events (the standard GP confidence event and the bias-envelope event) yields overall probability at least $1 - \delta$, and relabeling $B_{M,K,\delta/2}$ as $B_{M,K,\delta}$ gives (46). ■

Lemma 5 (Instantaneous regret with bias)) *On the event of Lemma 4, the UCB choice* $w_t \in \arg\max_w \mu_{t-1}(w) + \sqrt{\beta_t}\sigma_{t-1}(w)$ *satisfies*

$$r_t^{(g)} \; := \; g_M(w_M^\star) - g_M(w_t) \; \le \; 2\sqrt{\beta_t}\, \sigma_{t-1}(w_t) \; + \; 2\, B_{M,K,\delta}, \quad w_M^\star \in \arg\max_w g_M(w). \tag{59}$$

Proof Define the upper and lower confidence bounds (at round $t-1$) by

$$\mathrm{UCB}_{t-1}(w) \; := \; \mu_{t-1}(w) \; + \; \sqrt{\beta_t}\, \sigma_{t-1}(w), \qquad \mathrm{LCB}_{t-1}(w) \; := \; \mu_{t-1}(w) \; - \; \sqrt{\beta_t}\, \sigma_{t-1}(w). \tag{60}$$

By Lemma 4, on the stated event we have, simultaneously for all $w \in \mathcal{W}$,

$$g_M(w) \; \le \; \mathrm{UCB}_{t-1}(w) \; + \; B_{M,K,\delta}, \tag{61}$$

and

$$g_M(w) \; \ge \; \mathrm{LCB}_{t-1}(w) \; - \; B_{M,K,\delta}. \tag{62}$$

By the selection rule (59),

$$\mathrm{UCB}_{t-1}(w_M^\star) \; \le \; \mathrm{UCB}_{t-1}(w_t). \tag{63}$$

Start from the instantaneous regret definition and apply (61) at $w = w_M^\star$ and (62) at $w = w_t$:

$$\begin{aligned} r_t^{(g)} &= \; g_M(w_M^\star) \; - \; g_M(w_t) && (64) \\ &\le \; \big(\mathrm{UCB}_{t-1}(w_M^\star) + B_{M,K,\delta}\big) \; - \; \big(\mathrm{LCB}_{t-1}(w_t) - B_{M,K,\delta}\big) && (65) \\ &= \; \mathrm{UCB}_{t-1}(w_M^\star) \; - \; \mathrm{LCB}_{t-1}(w_t) \; + \; 2\, B_{M,K,\delta}. && (66) \end{aligned}$$

Use (63) to upper-bound the first two terms in (66):

$$\mathrm{UCB}_{t-1}(w_M^\star) \; - \; \mathrm{LCB}_{t-1}(w_t) \; \le \; \mathrm{UCB}_{t-1}(w_t) \; - \; \mathrm{LCB}_{t-1}(w_t). \tag{67}$$

By the definitions in (60),

$$\mathrm{UCB}_{t-1}(w_t) \; - \; \mathrm{LCB}_{t-1}(w_t) \; = \; 2\,\sqrt{\beta_t}\, \sigma_{t-1}(w_t). \tag{68}$$

Combine (66), (67), and (68) to obtain

$$r_t^{(g)} \; \le \; 2\,\sqrt{\beta_t}\, \sigma_{t-1}(w_t) \; + \; 2\, B_{M,K,\delta}, \tag{69}$$

which is precisely (59). ■

Lemma 6 (Information gain and variance sum) *For the queried sequence $A = (w_1, \dots, w_T)$,*

$$I(y_{1:T}; g_M) \;=\; \tfrac{1}{2} \log\det\big(I + R^{-2} K_A^{(\mathcal{W})}\big) \;=\; \tfrac{1}{2} \sum_{t=1}^{T} \log\big(1 + R^{-2} \sigma_{t-1}^2(w_t)\big),$$

and

$$\sum_{t=1}^{T} \sigma_{t-1}^2(w_t) \;\le\; \frac{2}{\log(1 + R^{-2})}\, \gamma_T. \tag{70}$$

Proof Let $A = (w_1, \dots, w_T)$ and denote by $K_A^{(\mathcal{W})} \in \mathbb{R}^{T \times T}$ the Gram matrix with entries $[K_A^{(\mathcal{W})}]_{ij} = k_{\mathcal{W}}(w_i, w_j)$. Under a zero-mean GP prior with kernel $k_{\mathcal{W}}$ and Gaussian likelihood with variance R^2, the joint (marginal) distribution of the observation vector $y_A := (y_1, \dots, y_T)^\top$ is

$$y_A \;\sim\; \mathcal{N}\big(0,\; K_A^{(\mathcal{W})} + R^2 I\big), \tag{71}$$

while the conditional distribution given the latent vector $g_{M,A} := (g_M(w_1), \dots, g_M(w_T))^\top$ is

$$y_A \,\big|\, g_M \;\sim\; \mathcal{N}\big(g_{M,A},\; R^2 I\big). \tag{72}$$

(a) Closed-form mutual information. By the definition of mutual information,

$$\begin{aligned} I(y_{1:T}; g_M) &= H(y_{1:T}) \;-\; H(y_{1:T} \mid g_M) && (73)\\ &= \tfrac{1}{2} \log\det\big(2\pi e\, (K_A^{(\mathcal{W})} + R^2 I)\big) \;-\; \tfrac{1}{2} \log\det\big(2\pi e\, R^2 I\big) && (74)\\ &= \tfrac{1}{2} \log\det\big(I + R^{-2} K_A^{(\mathcal{W})}\big). && (75) \end{aligned}$$

Equation (74) uses the Gaussian entropy formula $H(\mathcal{N}(m, \Sigma)) = \frac{1}{2} \log\big((2\pi e)^d \det \Sigma\big)$.

(b) Incremental identity. Let $\mathcal{F}_{t-1}$ be the sigma-field generated by $\{(w_\tau, y_\tau)\}_{\tau=1}^{t-1}$. The chain rule for mutual information gives

$$I(y_{1:T}; g_M) \;=\; \sum_{t=1}^{T} I\big(y_t;\; g_M \,\big|\, y_{1:t-1}\big). \tag{76}$$

Condition on $\mathcal{F}_{t-1}$. Under Gaussian process regression,

$$g_M(w_t) \,\big|\, \mathcal{F}_{t-1} \;\sim\; \mathcal{N}\big(\mu_{t-1}(w_t),\; \sigma_{t-1}^2(w_t)\big), \tag{77}$$

and, with independent Gaussian likelihood noise,

$$y_t \,\big|\, g_M,\; \mathcal{F}_{t-1} \;\sim\; \mathcal{N}\big(g_M(w_t),\; R^2\big). \tag{78}$$

Marginalizing (78) over (77) yields

$$y_t \,\big|\, \mathcal{F}_{t-1} \;\sim\; \mathcal{N}\big(\mu_{t-1}(w_t),\; \sigma_{t-1}^2(w_t) + R^2\big). \tag{79}$$

Therefore, by Gaussian entropies,

$$I\big(y_t;\ g_M \bigm| y_{1:t-1}\big) = H\big(y_t \bigm| y_{1:t-1}\big) - H\big(y_t \bigm| g_M,\ y_{1:t-1}\big) \tag{80}$$

$$= \tfrac{1}{2}\log\big(2\pi e\,(\sigma^2_{t-1}(w_t) + R^2)\big) - \tfrac{1}{2}\log\big(2\pi e\,R^2\big) \tag{81}$$

$$= \tfrac{1}{2}\log\Big(1 + R^{-2}\sigma^2_{t-1}(w_t)\Big). \tag{82}$$

Summing (82) over t and invoking (76) gives the incremental identity

$$I(y_{1:T}; g_M) \ = \ \tfrac{1}{2}\sum_{t=1}^{T}\log\Big(1 + R^{-2}\sigma^2_{t-1}(w_t)\Big). \tag{83}$$

Equalities (75) and (83) are known to coincide, which proves the first part of (70).

(c) Variance–sum bound. Assume the kernel is rescaled so that

$$k_W(w, w) \ \le \ 1 \qquad \text{for all } w \in \mathcal{W}. \tag{84}$$

Then GP posterior variances satisfy

$$0 \ \le \ \sigma^2_{t-1}(w_t) \ \le \ 1 \qquad \text{for all } t. \tag{85}$$

For any $u \in [0, R^{-2}]$, the concavity of $\log(1+u)$ implies the chord inequality

$$\log(1+u) \ \ge \ \frac{\log(1+R^{-2})}{R^{-2}}\,u. \tag{86}$$

Apply (86) with

$$u \ := \ R^{-2}\,\sigma^2_{t-1}(w_t) \ \in \ [0, R^{-2}] \qquad \text{(by (85))}, \tag{87}$$

to obtain, for each t,

$$\log\Big(1 + R^{-2}\sigma^2_{t-1}(w_t)\Big) \ \ge \ \log(1+R^{-2})\,\sigma^2_{t-1}(w_t). \tag{88}$$

Sum (88) over $t = 1, \dots, T$ and combine with (83):

$$\log(1+R^{-2})\sum_{t=1}^{T}\sigma^2_{t-1}(w_t) \ \le \ \sum_{t=1}^{T}\log\Big(1 + R^{-2}\sigma^2_{t-1}(w_t)\Big) \tag{89}$$

$$= \ 2\,I(y_{1:T}; g_M). \tag{90}$$

By the definition of the maximum information gain,

$$I(y_{1:T}; g_M) \ \le \ \gamma_T. \tag{91}$$

Therefore,

$$\sum_{t=1}^{T}\sigma^2_{t-1}(w_t) \ \le \ \frac{2}{\log(1+R^{-2})}\,\gamma_T, \tag{92}$$

which is (70). ■

Theorem 1 (Regret to the fixed-budget target g_M) *With probability at least $1-\delta$,*

$$\sum_{t=1}^{T}\big(g_M(w_M^\star)-g_M(w_t)\big) \;\le\; \sqrt{\frac{8}{\log(1+R^{-2})}\,T\,\beta_T\,\gamma_T} \;+\; 2T\,B_{M,K,\delta}. \tag{93}$$

Proof Let

$$r_t^{(g)} \;:=\; g_M(w_M^\star) \;-\; g_M(w_t), \qquad w_M^\star \in \arg\max_{w\in\mathcal{W}} g_M(w). \tag{94}$$

On the confidence event of Lemma 4, Lemma 5 gives, for every t,

$$r_t^{(g)} \;\le\; 2\sqrt{\beta_t}\,\sigma_{t-1}(w_t) \;+\; 2\,B_{M,K,\delta}. \tag{95}$$

Summing (95) over $t=1,\dots,T$ yields

$$\sum_{t=1}^{T} r_t^{(g)} \;\le\; 2\sum_{t=1}^{T}\sqrt{\beta_t}\,\sigma_{t-1}(w_t) \;+\; 2T\,B_{M,K,\delta}. \tag{96}$$

Apply Cauchy–Schwarz to the first term in (96):

$$\sum_{t=1}^{T}\sqrt{\beta_t}\,\sigma_{t-1}(w_t) \;\le\; \sqrt{\sum_{t=1}^{T}\beta_t}\,\sqrt{\sum_{t=1}^{T}\sigma_{t-1}^2(w_t)}. \tag{97}$$

Since β_t is nondecreasing in t,

$$\sum_{t=1}^{T}\beta_t \;\le\; T\,\beta_T. \tag{98}$$

By the variance–sum bound in Lemma 6,

$$\sum_{t=1}^{T}\sigma_{t-1}^2(w_t) \;\le\; \frac{2}{\log(1+R^{-2})}\,\gamma_T. \tag{99}$$

Combine (96), (97), (98), and (99):

$$\sum_{t=1}^{T} r_t^{(g)} \;\le\; 2\sqrt{T\,\beta_T}\,\sqrt{\frac{2}{\log(1+R^{-2})}\,\gamma_T} \;+\; 2T\,B_{M,K,\delta} \tag{100}$$

$$\;=\; \sqrt{\frac{8}{\log(1+R^{-2})}\,T\,\beta_T\,\gamma_T} \;+\; 2T\,B_{M,K,\delta}, \tag{101}$$

which is exactly (93). The bound holds on the confidence event of Lemma 4, which has probability at least $1-\delta$. ■

Theorem 2 (End-to-end regret to the original objective f) *Under PBRS anchoring and any uniform fixed-budget inner-loop accuracy bound $0 \le f(w)-g_M(w) \le \Delta_M$ for all $w\in\mathcal{W}$, with probability at least $1-\delta$,*

$$\sum_{t=1}^{T}\big(f(w^\star)-f(w_t)\big) \;\le\; \sqrt{\frac{8}{\log(1+R^{-2})}\,T\,\beta_T\,\gamma_T} \;+\; 2T\,\Delta_M \;+\; 2T\,B_{M,K,\delta}, \tag{102}$$

For the PPO inner loop under Assumptions A1–A6, Lemma 2 yields $\Delta_M = C_{\mathrm{ppo}}/\sqrt{M}$.

Proof Fix t and add/subtract $g_M(w^\star)$ and $g_M(w_t)$:

$$f(w^\star) - f(w_t) = \big(f(w^\star) - g_M(w^\star)\big) + \big(g_M(w^\star) - g_M(w_t)\big) + \big(g_M(w_t) - f(w_t)\big) \tag{103}$$

$$= \underbrace{r_t^{(g)}}_{\text{outer target gap}} + \underbrace{\big(f(w^\star) - g_M(w^\star)\big)}_{\text{inner gap at } w^\star} + \underbrace{\big(g_M(w_t) - f(w_t)\big)}_{\text{inner gap at } w_t}. \tag{104}$$

By the assumed uniform fixed-budget inner accuracy bound, we have

$$0 \le f(w) - g_M(w) \le \Delta_M \qquad \text{for all } w \in \mathcal{W}. \tag{105}$$

Apply (105) at $w^\star$ and w_t in (104) to obtain the per-round bound

$$f(w^\star) - f(w_t) \le r_t^{(g)} + 2\,\Delta_M. \tag{106}$$

Sum (106) over $t = 1, \dots, T$:

$$\sum_{t=1}^{T} \big(f(w^\star) - f(w_t)\big) \le \sum_{t=1}^{T} r_t^{(g)} + 2T\,\Delta_M. \tag{107}$$

Invoke Theorem 1 to bound the outer-target regret:

$$\sum_{t=1}^{T} r_t^{(g)} \le \sqrt{\frac{8}{\log(1+R^{-2})}\, T\,\beta_T\,\gamma_T} + 2T\,B_{M,K,\delta}. \tag{108}$$

Substitute (108) into (107):

$$\sum_{t=1}^{T} \big(f(w^\star) - f(w_t)\big) \le \sqrt{\frac{8}{\log(1+R^{-2})}\, T\,\beta_T\,\gamma_T} + 2T\,B_{M,K,\delta} + 2T\,\Delta_M \tag{109}$$

$$= \sqrt{\frac{8}{\log(1+R^{-2})}\, T\,\beta_T\,\gamma_T} + 2T\,\Delta_M + 2T\,B_{M,K,\delta}, \tag{110}$$

which is exactly (102). The probability statement follows from Theorem 1, which holds with probability at least $1-\delta$, and Lemma 2. ■