VALUE-DIRECTED EXPLORATION FOR EFFICIENT MODEL-BASED REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Efficient exploration remains a central challenge in reinforcement learning, particularly when environment interactions are costly. Curiosity-motivated algorithms guide exploration using model uncertainty as an intrinsic reward, however resolving errors in rarely visited or return-insensitive regions does not necessarily improve policy performance. Consequently, such strategies may allocate exploration effort inefficiently with respect to the task objective. To address this, we propose a value-directed optimistic exploration strategy that evaluates model uncertainty through its effect on future value, thereby prioritizing information acquisition in regions where reducing uncertainty can most affect which actions or policies are preferred. Our analysis yields a value-weighted information gain that can be substantially smaller than standard information gain when value sensitivity concentrates on a low-dimensional subset of the dynamics, giving regret bounds whose leading term tightens with this reduction and otherwise recovers value-agnostic scaling. Empirically, value-directed exploration improves performance in environments with task-irrelevant distractors while matching strong baselines on standard benchmarks.

1 INTRODUCTION

Reinforcement learning (RL) is a flexible paradigm for sequential decision-making in which agents learn policies through interactions with an unknown environment. While this has led to many successes in domains including robotics (Levine et al., 2016; Tang et al., 2025), games (Silver et al., 2017), and finetuning large language models (Ouyang et al., 2022), sample-efficient learning remains an active research problem. When data acquisition is costly, as is typical in real-world deployments, agents must learn from few interactions while balancing exploitation of the current policy with exploration of uncertain parts of the environment that may better inform future decisions. However, many widely used RL methods are highly sample-inefficient in practice (Schulman et al., 2017; Mnih et al., 2015).

Model-based reinforcement learning (MBRL) methods can improve sample-efficiency by estimating a dynamics model from experience and using it for planning or policy improvement, thereby allowing each observed transition to inform many future decisions (Janner et al., 2019; Hansen et al., 2024; Hafner et al., 2025). Yet most MBRL approaches rely on naive exploration strategies and do not consider how new data will affect future decisions. Optimism in the face of uncertainty is a principled strategy that selects actions according to favorable plausible outcomes (Brafman & Tennenholtz, 2002; Jaksch et al., 2010; Curi et al., 2020; Kakade et al., 2020). Classical optimistic methods often require expensive planning over confidence sets, motivating scalable variants to instead use model epistemic uncertainty to guide pure exploration (Sukhija et al., 2023) or as a bonus added to the task reward (Sukhija et al., 2025b). While

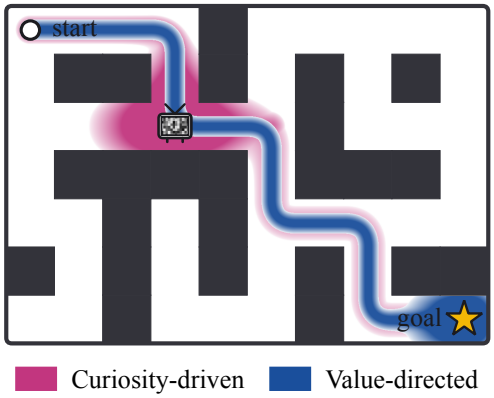


Figure 1: Curiosity-driven exploration is attracted to task-irrelevant model uncertainty around the noisy TV while value-directed exploration focuses on uncertainty relevant to reaching the goal.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

these approaches make optimism practical, their exploration signal is driven by model uncertainty rather than the uncertainty that matters for the control objective. This distinction is key because errors in rarely visited or value-insensitive states can contribute substantially to model uncertainty while being irrelevant to policy performance. A canonical example is the "noisy TV" problem, where curiosity-driven exploration can be drawn toward unpredictable but task-irrelevant dynamics, even when modeling those dynamics more accurately does not improve task performance (Burda et al., 2018; Mavor-Parker et al., 2022). Efficient exploration should therefore prioritize resolving model uncertainty according to its influence on future decisions and value.

To this end, we propose *Value-Directed Optimistic Model-Based Reinforcement Learning (VDO-MBRL)*, an exploration strategy that translates model uncertainty into its local relevance to future returns. We characterize this relevance through a first-order expansion of the task value, which shows that next-state prediction errors affect value through their alignment with the value gradient. Building on this insight, we develop two value-directed bonuses: a gradient-based approach that weights model uncertainty along value-sensitive directions and a gradient-free spread-based approximation that measures disagreement directly in predicted future values. In both cases, the bonus augments the extrinsic task reward to form an optimistic objective that favors exploration where model uncertainty can meaningfully affect task performance.

Contributions. Our contributions are threefold. First, we formulate a scalable optimistic exploration framework that focuses data acquisition on value-relevant model uncertainty, with practical gradient-based and gradient-free realizations. Second, we introduce a value-weighted information gain and derive an instance-dependent regret bound where the leading term tightens when model uncertainty concentrates in value-irrelevant directions. We further characterize when value relevance reduces the effective uncertainty governing exploration and identify structural conditions that allow for stronger improvements. Third, we empirically show that, unlike curiosity-motivated approaches, VDO-MBRL is robust to the noisy TV trap in maze environments and maintains competitive performance on DeepMind Control and Atari tasks.

2 Related Works

Optimistic exploration in MBRL. Optimism in the face of uncertainty is a classical principle for exploration across sequential decision-making problems (Auer et al., 2002; Srinivas et al., 2010; Jaksch et al., 2010). In MBRL, optimistic algorithms typically plan under favorable dynamics within a confidence set, achieving strong regret guarantees at the cost of joint optimization over policies and plausible models (Kakade et al., 2020). H-UCRL (Curi et al., 2020) enables more tractable optimistic planning by reparameterizing plausible dynamics with hallucinated controls over model uncertainty, however this enlarges the action space by the state dimension and remains expensive in high dimensions. Bonus-based methods provide another route towards tractability, with count and value dependent confidence bonuses in tabular MDPs avoiding joint optimization but relying on visitation counts or empirical tabular estimates (Strehl & Littman, 2008; Azar et al., 2017; Zanette & Brunskill, 2019). SOMBRL (Sukhija et al., 2025b) extends the bonus-based approach to continuous MDPs by planning under the learned mean dynamics with a model uncertainty bonus, achieving sublinear regret when the information gain grows sufficiently slowly. Complementary work estimates posterior value uncertainty through an uncertainty Bellman equation that is used for optimistic policy optimization (Luis et al., 2023) but the variance guarantees assume independent transition beliefs across states. VDO-MBRL retains the scalable planning structure of SOMBRL but constructs the exploration bonus from model uncertainty evaluated by its effect on future value, yielding a task-directed objective whose analysis can exploit task structure that is absent from global information gain.

Noisy TV problem. This problem reflects a broader challenge in real-world settings where dynamics unrelated to the objective can produce large prediction errors or model uncertainty and thereby attract curiosity-driven methods (Pathak et al., 2017; Burda et al., 2018). Prediction-error-based curiosity is particularly susceptible to this, as it assigns intrinsic reward to transitions that are difficult to predict, drawing exploration towards dynamic distractors (Pathak et al., 2017; Burda et al., 2018). Existing approaches primarily address this by distinguishing reducible from irreducible uncertainty. Disagreement-based and noise-aware methods aim to isolate epistemic uncertainty or suppress intrinsic rewards arising from stochastic dynamics (Pathak et al., 2019; Mavor-Parker et al., 2022; Jarrett

et al., 2023) while learning-progress methods reward improvements in prediction so that persistently unlearnable transitions cannot incentivize exploration (Kim et al., 2020; Hou et al., 2026). However, even reducible epistemic uncertainty can be irrelevant to control when improving the corresponding dynamics model does not improve the policy. VDO-MBRL instead evaluates model uncertainty through its effect on future value, making task relevance rather than reducibility the criterion for exploration. Figure 1 illustrates this distinction, with curiosity-driven exploration concentrating around the noisy TV and value-directed exploration focusing on regions relevant to reaching the goal.

Value-aware learning and exploration. Several works use value information beyond policy optimization to guide model learning and exploration. Value-aware model learning methods adapt the model objective during training to reflect how model errors affect downstream value (Farahmand et al., 2017), with practical methods using value or policy sensitivity to emphasize task-relevant state dimensions and visual features (Voelcker et al., 2022; Hutson et al., 2024). By incorporating control relevance into model learning, these approaches discern which aspects of the environment require accurate modeling and improve robustness when irrelevant state dimensions or predictable background dynamics compete for model capacity. QFLEX (Wei et al., 2026) addresses the more closely related exploration question of how to search efficiently in high-dimensional action spaces, using the learned critic to concentrate exploratory actions in high-value regions. VDO-MBRL instead uses value to determine which uncertainty in the dynamics model is worth resolving and should therefore drive information acquisition, rather than how the model is fit or exploratory actions are generated. This brings the classical value of information perspective, where information is valued through its effect on downstream decisions (Krause & Guestrin, 2009), to scalable optimistic exploration in MBRL.

3 Problem Setting

We consider a discrete-time dynamical system with state space $\mathcal{S} \subseteq \mathbb{R}^{d_\mathcal{S}}$ and action space $\mathcal{A} \subseteq \mathbb{R}^{d_\mathcal{A}}$. The system evolves according to

$$s_{t+1} = f^\star(s_t, a_t) + \epsilon_t, \qquad t = 0, \dots, T-1, \tag{1}$$

where $f^\star : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ is an unknown dynamics function, ϵ_t is process noise, and T is the episode length. Each episode starts from $s_0 \sim \rho_0$, where ρ_0 is a fixed distribution on $\mathcal{S}$. Given a reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, the return of a policy $\pi \in \Pi$ under dynamics f is

$$J(\pi, f) = \mathbb{E}_{\pi,f}\left[\sum_{t=0}^{T-1} r(s_t, a_t)\right] \quad \text{where} \quad s_{t+1} = f(s_t, a_t) + \epsilon_t, a_t \sim \pi(\cdot \mid s_t). \tag{2}$$

The optimal policy is one that maximizes the return i.e., $\pi^\star = \arg\max_{\pi \in \Pi} J(\pi, f^\star)$.

For the value-sensitivity analysis, we also define the finite-horizon state-value function

$$V_t^{\pi,f}(s) = \mathbb{E}_{\pi,f}\left[\sum_{\tau=t}^{T-1} r(s_\tau, a_\tau) \,\middle|\, s_t = s\right]. \tag{3}$$

This is the expected remaining return from state s at time t under the policy π and dynamics f.

At the start of each episode n, the learner selects and rolls out a policy π_n on the true system and observes $\mathcal{D}_n = \{(s_{n,t}, a_{n,t}, s_{n,t+1})\}_{t=0}^{T-1}$, which is used to update a probabilistic model of the unknown dynamics. We assume the reward function is known, as is standard in MBRL analyses (Curi et al., 2020; Sukhija et al., 2025b). If the reward is unknown, it can be learned jointly with the dynamics as an additional output and the analysis extends accordingly.

The goal is to learn a policy that performs as well as $\pi^\star$ as quickly as possible. We measure learning efficiency through the cumulative regret of the policies $\{\pi_n\}_{n=1}^N$ executed during learning,

$$R_N = \sum_{n=1}^{N} \big[J(\pi^\star, f^\star) - J(\pi_n, f^\star)\big]. \tag{4}$$

Sublinear regret implies that the average performance gap between the deployed policies and the optimal policy vanishes as the number of episodes N grows.

3.1 ASSUMPTIONS

Our theoretical analysis considers Gaussian Process (GP)-based dynamics models while our experiments estimate model uncertainty with neural network ensembles for scalability, as described in Section 6. To this end, we make standard regularity assumptions on the system, following prior optimistic MBRL analyses (Curi et al., 2020; Sukhija et al., 2025b).

Assumption 3.1 (Regularity)**.** *The reward is bounded and measurable, with $r : \mathcal{S} \times \mathcal{A} \to [0, R_{\max}]$, and the process noise is independent across time with $\epsilon_t \sim \mathcal{N}(0, \sigma^2 I)$ for $\sigma > 0$.*

We also assume that the unknown dynamics lie in an RKHS, as commonly done in GP bandit and kernelized MBRL analyses to control function complexity and obtain uniform confidence bounds (Srinivas et al., 2012; Chowdhury & Gopalan, 2017; Curi et al., 2020; Rothfuss et al., 2023).

Assumption 3.2 (RKHS dynamics)**.** *For each output coordinate $j \in \{1, \dots, d_{\mathcal{S}}\}$, the function $f_j^\star$ lies in a reproducing kernel Hilbert space $\mathcal{H}_k$ with a kernel k and satisfies $\|f_j^\star\|_k \leq B$. The kernel satisfies $k(z, z) \leq \sigma_{\max}$ for all $z = (s, a) \in \mathcal{S} \times \mathcal{A}$.*

Under Assumptions 3.1 and 3.2, the GP posterior uncertainty can be used to construct high-probability confidence intervals for the unknown dynamics. We use the following notion of calibration:

Definition 3.3 (All-time calibrated dynamics model (Rothfuss et al., 2023; Sukhija et al., 2025b))**.** *A probabilistic dynamics model $\{(\mu_n, \sigma_n, \beta_n)\}_{n \geq 0}$ is all-time calibrated if, with probability at least $1 - \delta$, $|f_j^\star(z) - \mu_{n,j}(z)| \leq \beta_n(\delta)\sigma_{n,j}(z)$ for all $z \in \mathcal{S} \times \mathcal{A}$, all coordinates $j \in \{1, \dots, d_{\mathcal{S}}\}$, and all episodes $n \geq 0$, where $\beta_n(\delta)$ is monotonically increasing in n.*

We assume the learner maintains calibrated posterior uncertainty throughout training as this is key for optimistic planning. For GPs, the calibration property follows from standard RKHS confidence bounds (Srinivas et al., 2012; Chowdhury & Gopalan, 2017). GPs are a natural model class for our analysis because they provide closed-form uncertainty estimates and can represent complex nonlinear functions through the choice of kernel (Rasmussen & Williams, 2006). Our framework can also use other uncertainty-aware model classes with calibrated uncertainty estimates (Curi et al., 2020).

Assumption 3.4 (Calibrated dynamics model)**.** *The learner maintains an all-time calibrated dynamics model in the sense of Definition 3.3.*

For a GP with Gaussian observation noise, let $\mathcal{Z}_{1:n}$ be the state-action inputs in $\mathcal{D}_{1:n}$. The posterior mean and variance for coordinate j are

$$\mu_{n,j}(z) = k_n(z)^\top (K_n + \sigma^2 I)^{-1} y_{1:n}^j, \qquad \sigma_{n,j}^2(z) = k(z, z) - k_n(z)^\top (K_n + \sigma^2 I)^{-1} k_n(z), \quad (5)$$

where $k_n(z) = [k(z, z_i)]_{z_i \in \mathcal{Z}_{1:n}}$, $K_n = [k(z_i, z_\ell)]_{z_i, z_\ell \in \mathcal{Z}_{1:n}}$, and $y_{1:n}^j$ denotes the noisy observations of the jth dynamics coordinate. We write $\sigma_n(z) = (\sigma_{n,1}(z), \dots, \sigma_{n,d_{\mathcal{S}}}(z))$.

3.2 BACKGROUND

Since the dynamics model is learned online, it can be inaccurate early in training and may remain uncertain in poorly explored regions, making the choice of which policy to execute next central to both the data collected and the accuracy of future policy evaluations. We therefore review policy selection strategies and the complexity measure underlying optimistic exploration analyses in MBRL.

Greedy planning. Greedy methods maximize the task return averaged over the dynamics posterior,

$$\pi_n^{\text{greedy}} = \arg\max_{\pi \in \Pi} \mathbb{E}_{f \sim p(f \mid \mathcal{D}_{1:n-1})}[J(\pi, f)]\,. \quad (6)$$

Such strategies are scalable and commonly used (Janner et al., 2019; Chua et al., 2018; Hafner et al., 2025), but they primarily exploit the current model and may avoid regions where model errors could change which policy is optimal, thus discouraging exploration.

Thompson sampling. This strategy draws a plausible dynamics model from the posterior distribution and plans under that sample,

$$f_n \sim p(f \mid \mathcal{D}_{1:n-1}), \qquad \pi_n^{\text{TS}} = \arg\max_{\pi \in \Pi} J(\pi, f_n). \quad (7)$$

Thompson sampling can induce exploration in a theoretically principled manner (Thompson, 1933; Russo et al., 2018), but requires posterior sampling over dynamics which can be challenging to scale.

Optimism. Optimistic exploration instead treats uncertain but plausible dynamics as potentially favorable. Seminal optimistic methods select a policy and dynamics model that jointly maximize the return over a confidence set, which yields principled exploration but is often computationally prohibitive (Curi et al., 2020; Kakade et al., 2020). Bonus-based optimistic methods provide a scalable relaxation by planning under the posterior mean dynamics while adding an uncertainty bonus to the reward. For instance, SOMBRL (Sukhija et al., 2025b) selects policies by solving

$$\pi_n^{\text{SOMBRL}} = \arg\max_{\pi \in \Pi} \mathbb{E}_{\pi, \mu_{n-1}} \left[\sum_{t=0}^{T-1} r(s_t, a_t) + \lambda_n \|\sigma_{n-1}(s_t, a_t)\| \right], \tag{8}$$

where $(\mu_{n-1}, \sigma_{n-1})$ is the GP posterior after $n-1$ episodes and $\lambda_n > 0$ balances exploration with exploitation. This objective encourages the agent to visit state-action pairs where the model is uncertain, thereby reducing epistemic uncertainty over time. The performance of such methods is governed by the maximum information gain of the kernel (Srinivas et al., 2012),

$$\Gamma_N(k) = \max_{A \subset \mathcal{S} \times \mathcal{A},\, |A| \leq NT} \frac{1}{2} \sum_{i=1}^{|A|} \log\bigl(1 + \sigma^{-2}\, \sigma_{i-1}^2(z_i)\bigr)\,, \tag{9}$$

which characterizes how quickly uncertainty can be reduced globally. Information gain bounds for several standard kernels are discussed in Appendix A.4.

Recent analyses of scalable, optimistic MBRL yield cumulative regret bounds of the form

$$R_N \leq \mathcal{O}\bigl(\Gamma_N^{3/2}(k)\sqrt{N}\bigr) \quad \text{(Theorem 5.4 of Sukhija et al. (2025b))}. \tag{10}$$

Kakade et al. (2020) derive a tighter bound of $\mathcal{O}(\Gamma_N(k)\,\sqrt{N})$, improving over equation 10 by a factor of $\sqrt{\Gamma_N(k)}$, but require joint optimization over policies and plausible dynamics which can be prohibitively expensive. In Appendix E, we sharpen equation 10 to match the N and information gain dependence of Kakade et al. (2020) using a technique for more tightly controlling accumulated model uncertainty that may be of independent interest. A key remaining limitation of these works is that $\Gamma_N(k)$ captures the global complexity of learning the dynamics across the entire state–action space without distinguishing which regions impact the control objective and obtainable value.

4 Value-Directed Optimistic Exploration

The utility of model uncertainty for exploration ultimately depends on how the corresponding model errors affect future decisions. Intuitively, model errors in value-sensitive regions can meaningfully change future return estimates and thereby alter action or policy selection, while errors in regions with low value variability across plausible next states are less likely to affect these decisions. This motivates distinguishing model uncertainty by its effect on future value rather than treating all uncertainty as equally useful for exploration. We operationalize this principle with *Value-Directed Optimistic Model-Based Reinforcement Learning (VDO-MBRL)*, which uses the mean-model planning structure of equation 8 but replaces the value-agnostic uncertainty bonus with a value-directed exploration bonus:

$$\pi_n^{\text{VDO-MBRL}} = \arg\max_{\pi \in \Pi} \mathbb{E}_{\pi, \mu_{n-1}} \left[\sum_{t=0}^{T-1} r(s_t, a_t) + b_{n,t}^{\pi}(s_t, a_t) \right]. \tag{11}$$

Value-directed bonus. In this section, we derive the bonus term $b_{n,t}^{\pi}$ to remain sensitive to how model errors affect value while preserving optimism when planning under the learned mean dynamics. Specifically, the resulting augmented mean-model return of each candidate policy should upper bound its true return. This requires controlling the error between the return predicted under the mean model $J(\pi, \mu_{n-1})$ and the return under the true dynamics $J(\pi, f^\star)$. For this purpose, we define

$$W_{n,t}^{\pi}(s) = V_t^{\pi, \mu_{n-1}}(s), \qquad \overline{W}_{n,t}^{\pi}(s) = \mathbb{E}_{\epsilon}[W_{n,t}^{\pi}(s + \epsilon)] \tag{12}$$

as the value of a candidate policy π under the mean model and its noise-averaged continuation respectively. The performance-difference identity (Kakade et al., 2020, Lemma B.1) then decomposes

the return error into the sum of one-step differences in the continuation value under the true and mean dynamics. For $z_t = (s_t, a_t)$, we have

$$J(\pi, f^\star) - J(\pi, \mu_{n-1}) = \mathbb{E}_{\pi, f^\star}\left[\sum_{t=0}^{T-1}\left(\overline{W}^{\pi}_{n,t+1}(f^\star(z_t)) - \overline{W}^{\pi}_{n,t+1}(\mu_{n-1}(z_t))\right)\right]. \quad (13)$$

Since Gaussian smoothing makes $\overline{W}^{\pi}_{n,t+1}$ differentiable, we control each one-step discrepancy by a Taylor expansion around $\mu_{n-1}(z)$. Writing $e_n(z) = f^\star(z) - \mu_{n-1}(z)$ and $H_t = R_{\max}(T-t)$ gives

$$\left|\overline{W}^{\pi}_{n,t+1}(f^\star(z)) - \overline{W}^{\pi}_{n,t+1}(\mu_{n-1}(z))\right| \le \left|\nabla \overline{W}^{\pi}_{n,t+1}(\mu_{n-1}(z))^\top e_n(z)\right| + \frac{H_{t+1}}{\sqrt{2}\sigma^2}\|e_n(z)\|_2^2. \quad (14)$$

The first-order term measures the alignment between the model prediction error and the value gradient. Applying the calibration bounds of Definition 3.3 enables bounding the left hand side above by

$$q^{\pi}_{n,t}(z) = \min\left\{H_{t+1}, \beta_{n-1}\sum_{j=1}^{d_{\mathcal{S}}}\sigma_{n-1,j}(z)\left|\partial_j \overline{W}^{\pi}_{n,t+1}(\mu_{n-1}(z))\right| + \frac{H_{t+1}}{\sqrt{2}\sigma^2}\beta_{n-1}^2\|\sigma_{n-1}(z)\|_2^2\right\},$$

where the first-order term scales each posterior width by the corresponding continuation-value sensitivity. Figure 4 in Appendix B illustrates how the first-order term downweights task-irrelevant uncertainty and approximates the ideal one-step optimistic bonus. The resulting $q^{\pi}_{n,t}$ bounds the absolute value of each one-step continuation discrepancy in equation 13, so summing these terms along true trajectories controls the return error. Since equation 11 evaluates the bonus under the mean model, Appendix D transfers this bound between models using the capped correction $v_n(z) = \min\{\beta_{n-1}^2\|\sigma_{n-1}(z)\|_2^2, \sigma^2\}$, a simulation bound, and Young's inequality to justify the full bonus

$$b^{\pi}_{n,t}(z) = \frac{3}{2}q^{\pi}_{n,t}(z) + \frac{3}{2}C_{\mathrm{B}}^{\mathrm{task}}v_n(z), \qquad C_{\mathrm{B}}^{\mathrm{task}} = \max\left\{\frac{2R_{\max}T^3}{\sigma^2}, 1\right\}. \quad (15)$$

This full bonus is needed for the theoretical optimism guarantee under calibrated GP uncertainty, while the practical version keeps only the gradient-weighted first-order term, as described in Section 6.

Solving the objective. Note that the bonus $b^{\pi}_{n,t}$ depends on the candidate policy through its task value $W^{\pi}_{n,t+1}$, making the reward of equation 11 policy-dependent. Following prior theoretical analyses of optimistic MBRL (Kakade et al., 2020; Sukhija et al., 2025b), we separate the regret analysis from the optimization error. In practice, we decouple the bonus from the candidate policy by estimating the task value of the current policy and fixing the resulting bonus for each policy update.

5 Theoretical Results

The bonus construction in Section 4 leads naturally to a regret analysis where model uncertainty is weighted by its effect on future task value, which we develop in this section. For the analysis, let $V_t^\star$ denote the Bellman-optimal value under the true dynamics and $\overline{V}_t^\star(s) = \mathbb{E}_\epsilon[V_t^\star(s+\epsilon)]$ be the noise-averaged continuation. We measure the relevance of uncertainty in coordinate j at z through

$$w^\star_{t+1,j}(z) = \left|\partial_j \overline{V}^\star_{t+1}(f^\star(z))\right|. \quad (16)$$

For a query history $\mathcal{H}_N = (z_{n,t})_{n,t}$, these sensitivities weight the posterior uncertainty in each dynamics coordinate, giving the maximum value-weighted information gain

$$\Gamma^\star_N(k) = \sup_{\mathcal{H}_N}\frac{1}{2}\sum_{n=1}^{N}\sum_{t=0}^{T-1}\sum_{j=1}^{d_{\mathcal{S}}}\log\left(1 + \frac{[w^\star_{t+1,j}(z_{n,t})\sigma_{n-1,j}(z_{n,t})]^2}{\sigma^2}\right). \quad (17)$$

Relative to standard information gain in equation 9, uncertainty contributes less where the optimal continuation value has low sensitivity to model prediction errors.

Theorem 5.1 (Value-directed regret)**.** *Fix $\delta \in (0,1)$ and suppose Assumptions 3.1, 3.2, and 3.4 hold on $\mathcal{S} = \mathbb{R}^{d_{\mathcal{S}}}$. Assume that $\pi^\star \in \Pi$ attains the unrestricted Bellman-optimal return and use the nondecreasing calibration multipliers $\beta_n = \beta_n(\delta/3) \ge 1$ for the GP posterior in equation 5. Let π_n solve equation 11 with the bonus defined in equation 15. Then, with probability at least $1-\delta$,*

$$R_N = \widetilde{\mathcal{O}}\left(\beta_{N-1}\sqrt{N\,\Gamma^\star_N(k)} + \beta_{N-1}^2\Gamma_N(k)\right). \quad (18)$$

The proof is given in Appendix D and Appendices B and C derive the same bounds for ideal optimistic synthesis, including for the LC^3 (Kakade et al., 2020) and H-UCRL (Curi et al., 2020) algorithms. Appendix A develops technical tools that support these analyses, which Appendix E also uses to sharpen the SOMBRL bound to $\widetilde{\mathcal{O}}(\beta_{N-1}\sqrt{N\Gamma_N(k)})$. For the same calibration coefficient, our leading term has the same scale when $\Gamma^\star_N(k)$ is on the same order as $\Gamma_N(k)$ and tightens as $\Gamma^\star_N(k)$ becomes smaller. Appendices F and G characterize when this reduction occurs, with a split bound over high and low sensitivity regions and a noisy TV example where standard information gain grows with task-irrelevant components while the value-weighted gain does not. This improvement governs the overall rate when the quadratic correction term is absorbed into the leading term, which holds under standard RKHS calibration $\beta_{N-1} = \widetilde{\mathcal{O}}(\sqrt{\Gamma_N(k)})$ when $\Gamma_N(k) = \widetilde{\mathcal{O}}(N^{1/3})$ and $\Gamma^\star_N(k) = \Omega(1)$.

Although the leading term depends on the value-weighted information gain, the calibration coefficient can also depend on global information gain. The general RKHS confidence bound of Chowdhury & Gopalan (2017, Theorem 2) gives $\beta_{N-1} = \widetilde{\mathcal{O}}(\sqrt{\Gamma_N(k)})$, yielding $R_N = \widetilde{\mathcal{O}}(\sqrt{N\Gamma_N(k)\Gamma^\star_N(k)} + \Gamma_N(k)^2)$. Appendix H shows that it is not possible to replace $\Gamma_N(k)$ with $\Gamma^\star_N(k)$ in the calibration coefficient for general kernels. However, under additional structure, tighter calibration bounds can improve this rate. In particular, applying Abbasi-Yadkori et al. (2011) to independent feature blocks can give $\beta_{N-1} = \widetilde{\mathcal{O}}(1)$ and hence $R_N = \widetilde{\mathcal{O}}(\sqrt{N\Gamma^\star_N(k)} + \Gamma_N(k))$. Appendix H also discusses extending batched algorithms that avoid information gain dependence to our setting (Li & Scarlett, 2022).

6 Experiments

In our experiments, we study the effectiveness of value-directed optimistic exploration across controlled distractor settings and standard control benchmarks. Our analysis predicts the greatest performance benefit when value-weighted information gain is smaller than global information gain, corresponding to problems where model uncertainty lies along directions with negligible influence on future value. We therefore evaluate VDO-MBRL in settings designed to introduce task-irrelevant uncertainty, specifically in state-based PointMaze (Fu et al., 2020) and vision-based MiniWorld (Chevalier-Boisvert et al., 2023) noisy TV environments, as well as standard benchmarks from the DeepMind Control Suite (Tassa et al., 2018) and Atari (Bellemare et al., 2013) where no such structure is imposed. For all learning curves, lines and shading show the mean and ± 1 standard error respectively. The full implementation details and hyperparameters are provided in Appendix I.

VDO-MBRL in practice. For scalability, we estimate model uncertainty with an ensemble of neural networks in place of the calibrated GP posterior used in our analysis, following prior works that transition from theory to practice (Curi et al., 2020; Sukhija et al., 2025b). For the ensemble approximation, let s'_f denote the next-state prediction of member f, $\bar{s}'$ their mean, and Σ their empirical covariance. We consider two practical variants: i) VDO-grad which follows the first-order task-directed bonus from Section 4 and omits the higher-order corrections and ii) VDO-spread which measures the spread of task values across ensemble-predicted next states. More precisely, letting V represent the learned value for the current policy, the two exploration bonuses are

$$b^{\text{grad}}(s,a) = \sqrt{\nabla V(\bar{s}')^\top \Sigma \nabla V(\bar{s}')}, \qquad b^{\text{spread}}(s,a) = \text{Std}_f\left[V(s'_f)\right]. \tag{19}$$

Note that VDO-grad is a first-order approximation to VDO-spread and the coordinatewise form $\sum_j |\partial_j V(\bar{s}')|\sigma_j$ we use for state-based control upper-bounds the same linearized value spread and directly matches the first-order term from Section 4. VDO-grad closely mirrors the theoretical first-order construction but relies on local value gradients and a linear approximation while VDO-spread requires only value evaluations and captures nonlinear variation across value predictions, giving a complementary gradient-free realization of the same principle. Therefore, we evaluate both variants to test whether value-directed exploration is robust to the choice of value sensitivity estimation.

Baselines. Our main exploration baseline is SOMBRL (Sukhija et al., 2025b), which augments the extrinsic task reward with a model uncertainty bonus following equation 8. For state-based control, we implement VDO-MBRL and SOMBRL on top of MBPO (Janner et al., 2019) and use DreamerV3 (Hafner et al., 2025) for their visual-control implementations. In all experiments, VDO-MBRL and SOMBRL use the same underlying learner and uncertainty estimator so that their

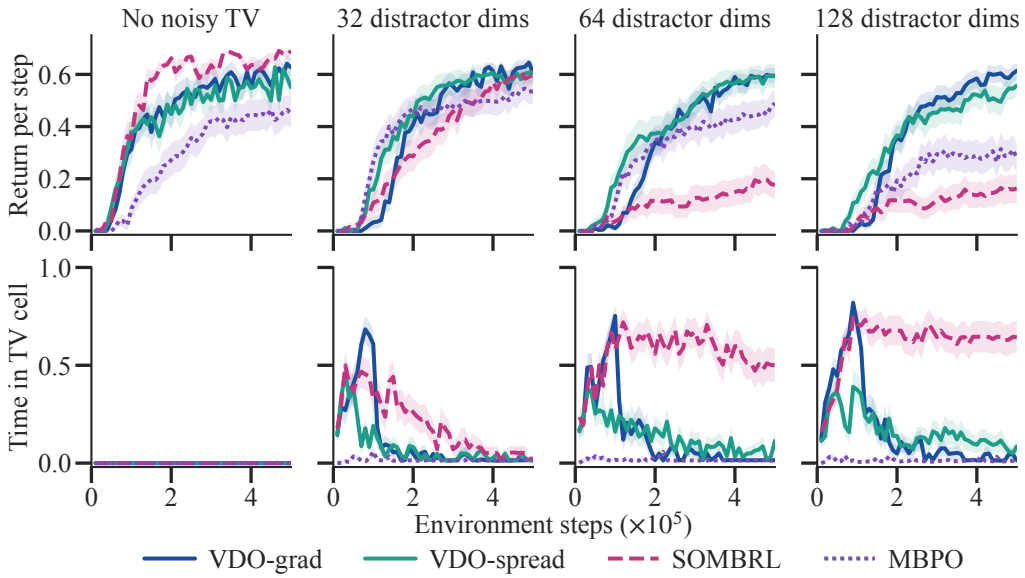


Figure 2: *Top:* Return per step over 25 seeds for PointMaze Large. *Bottom:* Corresponding curves for the amount of time spent near the noisy TV. As distractor dimensionality increases, both VDO-MBRL variants remain robust while SOMBRL spends more time at the TV and achieves lower returns.

comparison isolates the effect of value-directed exploration. We also include the unmodified backbone as a reference method without an explicit model uncertainty bonus.

State-based noisy TV. Inspired by prior work on the noisy TV failure mode of curiosity-driven exploration (Burda et al., 2018; Mavor-Parker et al., 2022), we consider a state-based noisy TV setting as a proof of concept for our central hypothesis that task-irrelevant uncertainty should not drive exploration. To this end, we use PointMaze with observations augmented by d distractor dimensions that are zero outside a TV cell and independently resampled from a Gaussian distribution when the agent is inside the cell. This introduces irreducible uncertainty that neither alters the dynamics nor provides task reward and creates a direct trade-off between exploring the TV and reaching the goal quickly to accumulate sparse rewards. Unlike prediction-error curiosity-driven methods, SOMBRL uses ensemble disagreement as the intrinsic reward, which can vanish even under irreducible transition noise, making this a more stringent noisy TV test. Figure 2 shows that as d grows, SOMBRL spends more time in the TV cell and achieves progressively lower returns, consistent with the distractor coordinates contributing increasingly to its uncertainty bonus. In contrast, both VDO-MBRL variants maintain lower TV occupancy rates and higher returns. MBPO also avoids the noisy TV trap because its exploration strategy is not driven by model uncertainty and its stochastic policy often provides sufficient exploration to solve this maze. We also consider four more challenging layouts using $d = 64$, with results given in Appendix J in Figure 5. In these environments, the stronger performance of VDO-MBRL relative to MBPO shows that its advantage is not simply due to suppressing exploration, but to preserving uncertainty-driven exploration while reducing attention to task-irrelevant uncertainty. Appendix J provides extra information about these experiments, including an ablation study on the TV placement in Figure 7.

Vision-based noisy TV. We extend the same noisy TV principle to visual control where task-irrelevant dynamics are observed through pixels rather than explicitly introduced as distractor coordinates. This setting tests whether value-directed exploration scales to high-dimensional visual inputs and reflects a common challenge in robotics where agents may observe dynamic elements such as changing displays, backgrounds, or moving objects whose evolution may be learnable or stochastic but do not affect how the objective should be achieved. To study both deterministic and stochastic visual distractors in a controlled setting, we augment the MiniWorld FourRooms environment with a screen near the start location and place the goal in a neighboring room. The agent observes first-person RGB images and can update the screen by pressing a button when nearby,

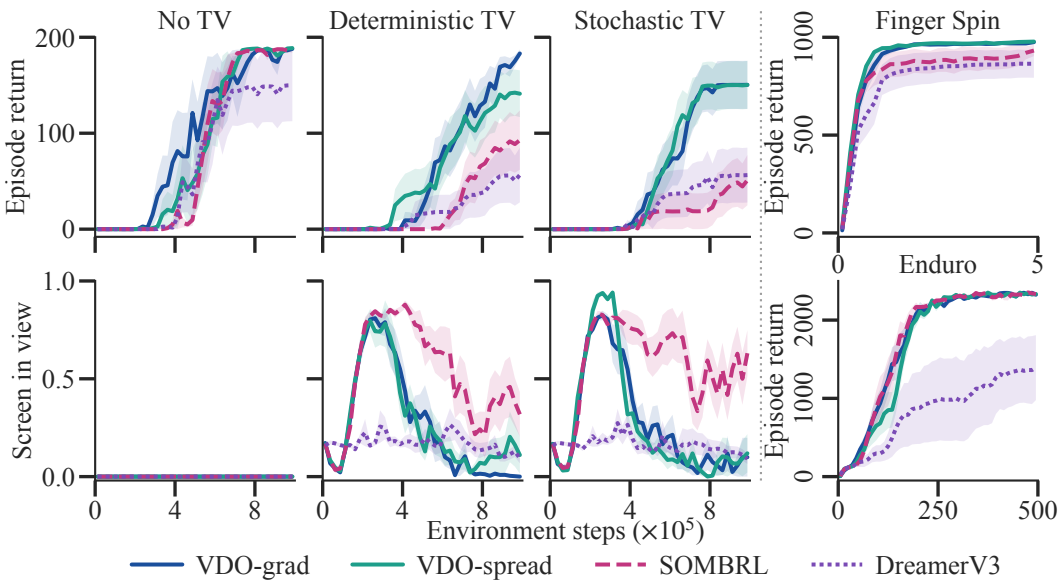


Figure 3: *Left:* The top row shows episode return over 10 seeds on MiniWorld FourRooms with an added TV screen and bottom gives the corresponding fraction of steps where the TV is visible. *Right:* Learning curves over 5 seeds for vision-based tasks from DMC and Atari. VDO-MBRL learns more efficiently than the baselines when a TV is present and otherwise closely tracks SOMBRL.

creating a choice between interacting with the TV and progressing towards the goal. We compare a deterministic TV whose 4×4 tile color pattern evolves according to a fixed deterministic rule, a stochastic TV where tiles are independently resampled after each update, and a control without a TV. Figure 3 shows that the methods perform similarly without a TV while both VDO variants learn most efficiently under the TV settings and spend less time with the screen in view than SOMBRL, with representative rollouts in Figure 9 illustrating this behavior. Their performance with the deterministic TV further demonstrates that value-directed exploration can ignore learnable but control-irrelevant uncertainty and thus addresses a broader class of distracting dynamics than irreducible noise alone.

Standard control benchmarks. Finally, we evaluate the VDO-MBRL variants on standard control tasks without deliberately introduced task-irrelevant uncertainty. Since our theory predicts the largest gains when model uncertainty contains a substantial value-irrelevant component, we expect smaller differences from SOMBRL on these benchmarks. Consistent with this prediction, the right side of Figure 3 as well as Figures 10 and 11 in Appendix J show broadly similar learning efficiency and final returns for VDO-MBRL and SOMBRL across visual DMC and Atari tasks and twelve state-based DMC tasks. Together with the noisy TV results, these findings support the conclusion that value-directed exploration is most beneficial when model uncertainty contains task-irrelevant components and shows no evidence of performance degradation on tasks without such structure.

7 Conclusion

We introduced VDO-MBRL, a scalable optimistic exploration strategy that focuses exploration on resolving model uncertainty with the greatest potential to affect future decisions. Our theoretical analysis formalizes this principle using a value-weighted information gain, giving instance-dependent regret bounds and characterizing when value relevance reduces exploration complexity and when additional structure enables stronger improvements. Empirically, the practical variants are robust to both stochastic and learnable distractors while remaining competitive on standard control benchmarks. Future work could extend these approaches on real robotic systems and study adaptively transitioning from broad initial exploration to value-directed exploration. Overall, this work points toward a more purposeful view of exploration where the goal is not to learn the environment uniformly but to efficiently acquire the information needed to make good decisions.

AI USE STATEMENT

In this work, we used generative AI tools to assist with the following: design the harder maze layouts shown in Figure 6, formulating and checking mathematical claims, providing ingredients for proving mathematical claims, writing proofs, providing feedback on research methodology, implementing methods, and interpreting results. We did not use generative AI tools to generate other synthetic datasets, develop theoretical models or conceptual frameworks, propose hypotheses, assist with translation, clean or reformat datasets, or support qualitative or thematic data analysis. Additionally, we used generative AI tools to create and modify scientific figures, suggest experimental parameters, create and edit code, draft and edits parts of the paper, and identify and understand relevant literature. We have carefully reviewed all AI-assisted work. For instance, all mathematical claims and proofs have been manually checked, code suggestions were inspected and tested, related works content was verified using the original papers, and figures were confirmed using the experiment results. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies general-purpose reinforcement learning methods using simulated benchmark environments and does not involve human subjects, personal data, or sensitive datasets. We are not aware of ethical concerns specific to this work beyond those associated with downstream deployment of reinforcement learning systems, which should be assessed in the relevant application context.

REPRODUCIBILITY STATEMENT

We state the assumptions underlying our theoretical analysis in Section 3 and provide complete proofs and supporting technical results in Appendices A–H. Section 6 describes the experimental settings, practical VDO-MBRL variants, baselines, and environment design with further implementation details in Appendix I, including the model architectures, hyperparameters, and computational resources used. Source code for VDO-MBRL, SOMBRL, and all experimental baselines used to produce the reported results is included in the supplementary materials.

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24, pp. 2312–2320, 2011.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256, 2002.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, 2017.

Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47: 253–279, 2013.

Ronen I Brafman and Moshe Tennenholtz. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3:213–231, 2002.

Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A Efros. Large-scale study of curiosity-driven learning. *arXiv preprint arXiv:1808.04355*, 2018.

Maxime Chevalier-Boisvert, Bolun Dai, Mark Towers, Rodrigo Perez-Vicente, Lucas Willems, Salem Lahlou, Suman Pal, Pablo Samuel Castro, and J K Terry. Minigrid & miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks. *Advances in Neural Information Processing Systems*, 36:73383–73394, 2023.

Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In *International Conference on Machine Learning*, pp. 844–853. PMLR, 2017.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Sebastian Curi, Felix Berkenkamp, and Andreas Krause. Efficient model-based reinforcement learning through optimistic policy search and planning. *Advances in Neural Information Processing Systems*, 33:14156–14170, 2020.

Amir-Massoud Farahmand, Andre Barreto, and Daniel Nikovski. Value-Aware Loss Function for Model-based Reinforcement Learning. In *International Conference on Artificial Intelligence and Statistics*, volume 54 of *PMLR*, pp. 1486–1494, 2017.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025.

Nicklas Hansen, Hao Su, and Xiaolong Wang. Td-mpc2: Scalable, robust world models for continuous control. *International Conference on Learning Representations*, 2024.

Zhibo Hou, Zhiyu An, and Wan Du. Beyond noisy-tvs: Noise-robust exploration via learning progress monitoring. In *International Conference on Learning Representations*, volume 2026, pp. 79369–79386, 2026.

Miles Hutson, Isaac Kauvar, and Nick Haber. Policy-shaped prediction: avoiding distractions in model-based reinforcement learning. *Advances in Neural Information Processing Systems*, 37: 13124–13148, 2024.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. In *Journal of Machine Learning Research*, volume 11, pp. 1563–1600, 2010.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

Daniel Jarrett, Corentin Tallec, Florent Altché, Thomas Mesnard, Rémi Munos, and Michal Valko. Curiosity in hindsight: Intrinsic exploration in stochastic environments. *International Conference on Machine Learning*, 2023.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *International Conference on Machine Learning*, pp. 267–274, 2002.

Sham Kakade, Akshay Krishnamurthy, Kendall Lowrey, Motoya Ohnishi, and Wen Sun. Information theoretic regret bounds for online nonlinear control. *Advances in Neural Information Processing Systems*, 33:15312–15325, 2020.

Kuno Kim, Megumi Sano, Julian De Freitas, Nick Haber, and Daniel Yamins. Active world model learning with progress curiosity. In *International Conference on Machine Learning*, pp. 5306–5315. PMLR, 2020.

Ilya Kostrikov. JAXRL: Implementations of Reinforcement Learning algorithms in JAX, 10 2021. URL `https://github.com/ikostrikov/jaxrl`.

Andreas Krause and Carlos Guestrin. Optimal value of information in graphical models. *Journal of Artificial Intelligence Research*, 35:557–591, 2009.

Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17(39):1–40, 2016.

Zihan Li and Jonathan Scarlett. Gaussian process bandit optimization with few batches. In *International Conference on Artificial Intelligence and Statistics*, volume 151, pp. 92–107. PMLR, 2022.

Carlos E. Luis, Alessandro G. Bottero, Julia Vinogradska, Felix Berkenkamp, and Jan Peters. Model-based uncertainty in value functions. In *International Conference on Artificial Intelligence and Statistics*, volume 206, 2023.

Augustine Mavor-Parker, Kimberly Young, Caswell Barry, and Lewis Griffin. How to stay curious while avoiding noisy tvs using aleatoric uncertainty estimation. In *International Conference on Machine Learning*, pp. 15220–15240. PMLR, 2022.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Ian Osband, John Aslanides, and Albin Cassirer. Randomized prior functions for deep reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International Conference on Machine Learning*, pp. 2778–2787. PMLR, 2017.

Deepak Pathak, Dhiraj Gandhi, and Abhinav Gupta. Self-supervised exploration via disagreement. In *International Conference on Machine Learning*, pp. 5062–5071. PMLR, 2019.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Jonas Rothfuss, Bhavya Sukhija, Tobias Birchler, Parnian Kassraie, and Andreas Krause. Hallucinated adversarial control for conservative offline policy evaluation. In *Uncertainty in Artificial Intelligence*, pp. 1774–1784. PMLR, 2023.

Daniel Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, and Zheng Wen. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–99, 2018.

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

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, 2017.

Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: no regret and experimental design. In *International Conference on Machine Learning*, pp. 1015–1022, 2010.

Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias W Seeger. Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory*, 58(5):3250–3265, 2012.

Alexander L. Strehl and Michael L. Littman. An analysis of model-based interval estimation for markov decision processes. *Journal of Computer and System Sciences*, 2008.

Bhavya Sukhija, Lenart Treven, Cansu Sancaktar, Sebastian Blaes, Stelian Coros, and Andreas Krause. Optimistic active exploration of dynamical systems. In *Advances in Neural Information Processing Systems*, volume 36, pp. 38122–38153, 2023.

Bhavya Sukhija, Stelian Coros, Andreas Krause, Pieter Abbeel, and Carmelo Sferrazza. Maxinforl: Boosting exploration in reinforcement learning through information gain maximization. *International Conference on Learning Representations*, 2025a.

Bhavya Sukhija, Lenart Treven, Carmelo Sferrazza, Florian Dörfler, Pieter Abbeel, and Andreas Krause. Sombrl: Scalable and optimistic model-based rl. In *Advances in Neural Information Processing Systems*, 2025b.

Chen Tang, Ben Abbatematteo, Jiaheng Hu, Rohan Chandra, Roberto Martín-Martín, and Peter Stone. Deep reinforcement learning for robotics: A survey of real-world successes. *Annual Review of Control, Robotics, and Autonomous Systems*, 8(1):153–188, 2025.

Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018.

William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

Luca Vignola, Bruce D. Lee, Manish Prajapat, Manuel Wendl, Melanie Zeilinger, Andreas Krause, and Yarden As. Sampling-based safe reinforcement learning. *arXiv preprint arXiv:2605.19469*, 2026.

Claas Voelcker, Victor Liao, Animesh Garg, and Amir-Massoud Farahmand. Value gradient weighted model-based reinforcement learning. *International Conference on Learning Representations*, 2022.

Yunyue Wei, Chenhui Zuo, and Yanan Sui. Scalable exploration for high-dimensional continuous control via value-guided flow. *International Conference on Learning Representations*, 2026.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, 2019.

Fedor Zhdanov and Yuri Kalnishkan. An identity for kernel ridge regression. *arXiv preprint arXiv:1112.1390*, 2011.

APPENDIX CONTENTS

A Technical Results **15**
A.1 Auxiliary-reward notation 15
A.2 Gaussian-smoothed continuations 15
A.3 Performance-difference and simulation bounds 17
A.4 Information-gain and concentration bounds 19
A.5 Accumulated GP prediction error 21
A.6 Common setting for the regret analyses 24

B Optimistic Synthesis and Hallucinated Control **24**
B.1 Exact optimistic synthesis 24
B.2 Optimistic exploration bonuses 24

C Self-Consistent Taylor Optimism **25**
C.1 Taylor elimination of hallucinated control 26
C.2 Policywise Taylor certificate and synthesis criterion 26
C.3 Optimal-value relevance 27
C.4 Regret bound 29

D Task-Only Value-Directed Exploration **30**
D.1 Task-only policy criterion 31
D.2 Comparing task-only and optimal-value sensitivity 32
D.3 Regret bound 32

E Undirected Confidence Bonuses: A Tightened SOMBRL Analysis **34**
E.1 Planning criterion and return transfers 34
E.2 Tightened regret bound 34

F When Value-Weighted Information Gain Is Smaller: A Split Bound **35**
F.1 Regional information gain and the split bound 35
F.2 Intuition and comparison with unweighted information gain 36

G A Noisy TV with Many Channels **37**

H Can Calibration Depend on the Value-Weighted Information Gain? **39**
H.1 The calibration question 39
H.2 A lower bound with separated information gains 40

I Experiment Details **44**

J Additional Experiments **47**

A TECHNICAL RESULTS

A.1 AUXILIARY-REWARD NOTATION

For a reward sequence $q = (q_t)_{t=0}^{T-1}$, define

$$J(\pi, f; q) \triangleq \mathbb{E}_{\pi,f}\left[\sum_{t=0}^{T-1} q_t(s_t, a_t)\right], \qquad V_{q,t}^{\pi,f}(s) \triangleq \mathbb{E}_{\pi,f}\left[\sum_{\tau=t}^{T-1} q_\tau(s_\tau, a_\tau) \,\middle|\, s_t = s\right]. \tag{20}$$

All returns use the same initial-state distribution $s_0 \sim \rho_0$ from equation 2, with a point mass for a fixed start. A stationary reward means $q_t = q$ and $V_{q,T}^{\pi,f} \equiv 0$. The task-return notation is $J(\pi, f) = J(\pi, f; r)$ and $V_t^{\pi,f} = V_{r,t}^{\pi,f}$.

A.2 GAUSSIAN-SMOOTHED CONTINUATIONS

Assume $\sigma > 0$. For a bounded measurable function $V : \mathbb{R}^{d_S} \to \mathbb{R}$ and candidate dynamics f, write

$$\overline{V}(y) \triangleq \mathbb{E}_\epsilon[V(y + \epsilon)], \qquad (P_f V)(z) \triangleq \overline{V}(f(z)), \qquad \epsilon \sim \mathcal{N}(0, \sigma^2 I). \tag{21}$$

Lemma A.1 (Gaussian smoothing and Taylor certificate)**.** *Let $V : \mathbb{R}^{d_S} \to [0, H]$ be bounded and measurable. Its Gaussian smoothing is infinitely differentiable and, for every coordinate j, satisfies*

$$\partial_j \overline{V}(y) = \frac{1}{\sigma^2}\mathbb{E}_\epsilon[\epsilon_j V(y + \epsilon)], \qquad \left|\partial_j \overline{V}(y)\right| \le \frac{c_g H}{\sigma}, \tag{22}$$

$$\|\nabla \overline{V}(y)\|_2 \le \frac{c_g H}{\sigma}, \qquad \left\|\nabla^2 \overline{V}(y)\right\|_{\mathrm{op}} \le \frac{c_H H}{\sigma^2}, \quad c_g \triangleq 1, \qquad c_H \triangleq \sqrt{2}. \tag{23}$$

Consequently, for all $x, y \in \mathbb{R}^{d_S}$,

$$\left|\overline{V}(x) - \overline{V}(y)\right| \le \sum_{j=1}^{d_S} \left|\partial_j \overline{V}(y)\right| |x_j - y_j| + \frac{c_H H}{2\sigma^2} \|x - y\|_2^2 . \tag{24}$$

If $|x_j - y_j| \le \beta s_j$ for a vector $s \in \mathbb{R}_+^{d_S}$, then the left-hand side is also at most

$$\min\left\{ H, \beta \sum_{j=1}^{d_S} |\partial_j \overline{V}(y)| s_j + \frac{c_H H}{2\sigma^2}\beta^2 \|s\|_2^2 \right\}. \tag{25}$$

Proof. Let $p_\sigma(u) = (2\pi\sigma^2)^{-d_S/2} \exp(-\|u\|_2^2/(2\sigma^2))$ be the Gaussian density. Changing variables $x = y + \epsilon$ gives

$$\overline{V}(y) = \int_{\mathbb{R}^{d_S}} V(x) p_\sigma(x - y)\, dx.$$

Here y appears only in the density, whose derivatives are

$$\nabla_y p_\sigma(x - y) = \frac{x - y}{\sigma^2} p_\sigma(x - y),$$
$$\nabla_y^2 p_\sigma(x - y) = \left(\frac{(x - y)(x - y)^\top}{\sigma^4} - \frac{I}{\sigma^2}\right) p_\sigma(x - y).$$

On any bounded neighborhood of y, each density derivative is bounded uniformly by a polynomial times a Gaussian tail, integrable in x. Since V is bounded, dominated differentiation applies at every order, so $\overline{V}$ is infinitely differentiable. Differentiating the integral and substituting $\epsilon = x - y$ back in gives

$$\nabla \overline{V}(y) = \int V(x) \frac{x - y}{\sigma^2} p_\sigma(x - y)\, dx = \frac{1}{\sigma^2}\mathbb{E}_\epsilon[\epsilon V(y + \epsilon)],$$
$$\nabla^2 \overline{V}(y) = \int V(x) \left(\frac{(x - y)(x - y)^\top}{\sigma^4} - \frac{I}{\sigma^2}\right) p_\sigma(x - y)\, dx$$
$$= \frac{1}{\sigma^2}\mathbb{E}_\epsilon\left[\left(\frac{\epsilon\epsilon^\top}{\sigma^2} - I\right) V(y + \epsilon)\right].$$

For every unit vector v, put $g = \sigma^{-1} v^\top \epsilon \sim \mathcal{N}(0,1)$. Cauchy–Schwarz and $V^2 \le H^2$ give

$$|v^\top \nabla \overline{V}(y)| = \frac{1}{\sigma} |\mathbb{E}[gV(y+\epsilon)]| \le \frac{1}{\sigma} \sqrt{\mathbb{E}[g^2]\,\mathbb{E}[V(y+\epsilon)^2]} \le \frac{H}{\sigma},$$

$$\begin{aligned} \left|v^\top \nabla^2 \overline{V}(y) v\right| &= \frac{1}{\sigma^2} \left|\mathbb{E}[(g^2-1)V(y+\epsilon)]\right| \\ &\le \frac{1}{\sigma^2} \sqrt{\mathbb{E}[(g^2-1)^2]\,\mathbb{E}[V(y+\epsilon)^2]} \le \frac{\sqrt{2}H}{\sigma^2}, \end{aligned}$$

using $\mathbb{E}[g^2] = 1$ and $\mathbb{E}[(g^2-1)^2] = 3 - 2 + 1 = 2$. Taking suprema over unit v proves the gradient and Hessian norm bounds. Taylor's theorem gives equation 24; substituting the coordinatewise radii and using $0 \le \overline{V} \le H$ gives equation 25. □

Lemma A.2 (Comparison of smoothed gradients)**.** *Let $U, V : \mathbb{R}^{d_S} \to [0, H]$ be bounded and measurable. For all $x, y \in \mathbb{R}^{d_S}$ and every coordinate j,*

$$\left| \, |\partial_j \overline{U}(x)| - |\partial_j \overline{V}(y)| \, \right| \le \frac{c_H H}{\sigma^2} \|x - y\|_2 + \frac{1}{\sigma} \sqrt{H\, \mathbb{E}_\epsilon |U(y+\epsilon) - V(y+\epsilon)|}. \tag{26}$$

Proof. By equation 23, transporting $\partial_j \overline{U}$ from x to y costs at most $c_H H \|x - y\|_2 / \sigma^2$. At y, the gradient identity equation 22 and Cauchy–Schwarz give

$$\begin{aligned} |\partial_j \overline{U}(y) - \partial_j \overline{V}(y)| &\le \frac{1}{\sigma^2} \sqrt{\mathbb{E}[\epsilon_j^2]} \sqrt{\mathbb{E}[(U(y+\epsilon) - V(y+\epsilon))^2]} \\ &\le \frac{1}{\sigma} \sqrt{H\, \mathbb{E}|U(y+\epsilon) - V(y+\epsilon)|}, \end{aligned}$$

because $|U - V| \le H$. The reverse triangle inequality completes the proof. □

Lemma A.3 (Gaussian shift controlled by endpoint second moments)**.** *Let $x, y \in \mathbb{R}^{d_S}$, let $v : \mathbb{R}^{d_S} \to \mathbb{R}_+$ be bounded and measurable, and let $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$. Then*

$$|\mathbb{E}_\epsilon[v(x+\epsilon)] - \mathbb{E}_\epsilon[v(y+\epsilon)]| \le \max_{h \in \{x,y\}} \sqrt{\mathbb{E}_\epsilon[v(h+\epsilon)^2]} \min\left\{ \frac{\|x-y\|_2}{\sigma}, 1 \right\}. \tag{27}$$

Proof. Lemma C.2 of Kakade et al. (2020) gives the one-sided inequality

$$\mathbb{E}_\epsilon[v(x+\epsilon)] - \mathbb{E}_\epsilon[v(y+\epsilon)] \le \sqrt{\mathbb{E}_\epsilon[v(x+\epsilon)^2]} \min\left\{ \frac{\|x-y\|_2}{\sigma}, 1 \right\}.$$

Exchanging x and y controls the opposite sign; taking the larger endpoint second moment proves equation 27. □

Corollary A.4 (Exact, Taylor, and undirected confidence widths)**.** *Let $V : \mathbb{R}^{d_S} \to [0, H]$ be bounded and measurable, fix $y \in \mathbb{R}^{d_S}$ and $\rho \in \mathbb{R}_+^{d_S}$, and define*

$$\begin{aligned} \Omega_\rho^+[V](y) &\triangleq \sup_{|\delta_j| \le \rho_j} \{\overline{V}(y+\delta) - \overline{V}(y)\}, \\ \Omega_\rho^\pm[V](y) &\triangleq \sup_{|\delta_j| \le \rho_j} |\overline{V}(y+\delta) - \overline{V}(y)|, \\ L_\rho[V](y) &\triangleq \sum_{j=1}^{d_S} \rho_j |\partial_j \overline{V}(y)|, \qquad\qquad Q_\rho(H) \triangleq \frac{c_H H}{2\sigma^2} \|\rho\|_2^2. \end{aligned}$$

Then

$$|\Omega_\rho^+[V](y) - L_\rho[V](y)| \le Q_\rho(H), \tag{28}$$

$$\Omega_\rho^\pm[V](y) \le \min\{H, L_\rho[V](y) + Q_\rho(H)\}, \tag{29}$$

$$L_\rho[V](y) \le \frac{c_g H}{\sigma} \|\rho\|_2, \tag{30}$$

$$\Omega_\rho^\pm[V](y) \le H \min\left\{ \frac{\|\rho\|_2}{\sigma}, 1 \right\}. \tag{31}$$

The Taylor bound retains local value sensitivity.

Proof. Taylor's theorem bounds the difference between $\overline{V}(y+\delta) - \overline{V}(y)$ and $\nabla\overline{V}(y)^\top\delta$ by $Q_\rho(H)$ uniformly over the confidence box, proving equation 28; the absolute-value version proves equation 29. Cauchy–Schwarz and the gradient-norm bound in Lemma A.1 prove equation 30. Lemma A.3, with both endpoint second moments bounded by H^2, gives equation 31. □

A.3 Performance-difference and simulation bounds

Lemma A.5 (Statewise performance-difference identity)**.** *Suppose $\sigma > 0$ and $\mathcal{S} = \mathbb{R}^{d_\mathcal{S}}$. Let $q_t : \mathcal{S}\times\mathcal{A} \to [0, q_{\max}]$ for $t = 0, \dots, T-1$, let $f, g : \mathcal{S}\times\mathcal{A} \to \mathcal{S}$, and fix $\pi \in \Pi$. For every start time τ and state s,*

$$V^{\pi,f}_{q,\tau}(s) - V^{\pi,g}_{q,\tau}(s) = \mathbb{E}_{\pi,f}\left[\sum_{t=\tau}^{T-1}\left((P_f V^{\pi,g}_{q,t+1})(z_t) - (P_g V^{\pi,g}_{q,t+1})(z_t)\right) \,\middle|\, s_\tau = s\right]. \tag{32}$$

Proof. Apply the Bellman recursion for $V^{\pi,g}_{q,t}$ along a trajectory generated by f and sum from $t = \tau$ to $T-1$. The value terms telescope and $V^{\pi,g}_{q,T} = 0$, which proves equation 32 (Kakade & Langford, 2002; Sukhija et al., 2023). □

Corollary A.6 (Statewise one-step certificate)**.** *Under the conditions of Lemma A.5, suppose nonnegative functions b_t satisfy*

$$\left|(P_f V^{\pi,g}_{q,t+1})(z) - (P_g V^{\pi,g}_{q,t+1})(z)\right| \le b_t(z)$$

for every t and z. Then, for every τ and s,

$$\left|V^{\pi,f}_{q,\tau}(s) - V^{\pi,g}_{q,\tau}(s)\right| \le \mathbb{E}_{\pi,f}\left[\sum_{t=\tau}^{T-1} b_t(z_t) \,\middle|\, s_\tau = s\right]. \tag{33}$$

Proof. Take absolute values in equation 32 and use the assumed one-step bounds. □

Corollary A.7 (Linear simulation bound)**.** *Under the conditions of Lemma A.5, for either $h = f$ or $h = g$,*

$$|J(\pi, f; q) - J(\pi, g; q)| \le \frac{q_{\max} T}{\sigma} J(\pi, h; \|f - g\|_2). \tag{34}$$

Proof. Since $0 \le V^{\pi,g}_{q,t+1} \le q_{\max} T$, Lemma A.3 gives

$$|(P_f - P_g)V^{\pi,g}_{q,t+1}(z)| \le \frac{q_{\max} T}{\sigma}\|f(z) - g(z)\|_2.$$

Average equation 32 at $\tau = 0$ over ρ_0 to obtain the case $h = f$; exchange f and g for the other case. □

Lemma A.8 (Accumulated continuation second moments)**.** *Under the conditions of Lemma A.5, define*

$$\begin{aligned} d_t(z) &\triangleq (P_f V^{\pi,g}_{q,t+1})(z) - (P_g V^{\pi,g}_{q,t+1})(z), \\ m_t(z) &\triangleq \max\left\{(P_f[(V^{\pi,g}_{q,t+1})^2])(z), (P_g[(V^{\pi,g}_{q,t+1})^2])(z)\right\}, \\ D_{f,g} &\triangleq \mathbb{E}_{\pi,f}\left[\sum_{t=0}^{T-1} |d_t(z_t)|\right]. \end{aligned}$$

Then

$$\mathbb{E}_{\pi,f}\left[\sum_{t=0}^{T-1} m_t(z_t)\right] \le 2q_{\max} T^2 \left(J(\pi, f; q) + D_{f,g}\right). \tag{35}$$

Proof. Because $0 \le V^{\pi,g}_{q,t+1} \le q_{\max} T$,

$$m_t(z) \le q_{\max} T \left((P_f V^{\pi,g}_{q,t+1})(z) + (P_g V^{\pi,g}_{q,t+1})(z) \right).$$

The tower property and the Bellman recursion give, respectively,

$$\mathbb{E}_{\pi,f}\left[\sum_{t=0}^{T-1} (P_f V^{\pi,g}_{q,t+1})(z_t)\right] = \mathbb{E}_{\pi,f}\left[\sum_{t=0}^{T-1} V^{\pi,g}_{q,t}(s_t)\right] - J(\pi, g; q),$$
$$\mathbb{E}_{\pi,f}\left[\sum_{t=0}^{T-1} (P_g V^{\pi,g}_{q,t+1})(z_t)\right] = \mathbb{E}_{\pi,f}\left[\sum_{t=0}^{T-1} V^{\pi,g}_{q,t}(s_t)\right] - J(\pi, f; q).$$

The first identity uses $V^{\pi,g}_{q,T} = 0$ and the common initial-state distribution. Both subtracted returns are nonnegative, so

$$\mathbb{E}_{\pi,f}\left[\sum_{t=0}^{T-1} m_t(z_t)\right] \le 2 q_{\max} T \, \mathbb{E}_{\pi,f}\left[\sum_{t=0}^{T-1} V^{\pi,g}_{q,t}(s_t)\right].$$

By equation 32,

$$V^{\pi,g}_{q,t}(s) \le \mathbb{E}_{\pi,f}\left[\sum_{\ell=t}^{T-1} (q_\ell + |d_\ell|)(z_\ell) \,\middle|\, s_t = s\right].$$

Taking expectations and summing over t counts each reward and error at most T times. Thus the expected value sum is at most $T(J(\pi, f; q) + D_{f,g})$, proving the claim. □

Lemma A.9 (Value-weighted simulation bound)**.** *Under the conditions of Lemma A.5, define the clipped squared model error $e^2_{f,g}(z) \triangleq \min\{\|f(z) - g(z)\|_2^2, \sigma^2\}$. For either $h = f$ or $h = g$,*

$$|J(\pi, f; q) - J(\pi, g; q)| \le C_\Delta J(\pi, h; e^2_{f,g}) + \sqrt{C_\Delta J(\pi, h; e^2_{f,g}) J(\pi, h; q)}, \tag{36}$$

where

$$C_\Delta = \max\left\{ \frac{2 q_{\max} T^2}{\sigma^2}, 1 \right\}. \tag{37}$$

Replacing $e^2_{f,g}$ by the larger reward $\|f - g\|_2^2$ gives the untruncated form.

Proof. Take $h = f$ and use d_t, m_t, and $D_{f,g}$ from Lemma A.8. Lemma A.5 and the triangle inequality give

$$|J(\pi, f; q) - J(\pi, g; q)| \le D_{f,g}.$$

Lemma A.3, followed by Cauchy–Schwarz, gives

$$D_{f,g} \le \frac{1}{\sigma} \sqrt{\mathbb{E}_{\pi,f}\left[\sum_{t=0}^{T-1} m_t(z_t)\right]} \sqrt{J(\pi, f; e^2_{f,g})}.$$

Combining this with equation 35, and writing $K = C_\Delta J(\pi, f; e^2_{f,g})$, yields

$$D^2_{f,g} \le K \left(J(\pi, f; q) + D_{f,g} \right).$$

Solving this scalar quadratic and using $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ gives

$$D_{f,g} \le K + \sqrt{K J(\pi, f; q)},$$

which proves equation 36 for $h = f$. Exchanging f and g proves the case $h = g$. □

A.4 INFORMATION-GAIN AND CONCENTRATION BOUNDS

Bounded features on an unbounded domain. For $\sigma > 0$, the full support of Gaussian transition noise requires $\mathcal{S} = \mathbb{R}^{d_\mathcal{S}}$, whereas standard information-gain rates for squared-exponential and Matérn kernels are stated on compact domains. Following the construction in Vignola et al. (2026, Appendix A.1), let $\mathcal{Z} = \mathcal{S} \times \mathcal{A}$ and use $k(z, z') = \widetilde{k}(\psi(z), \psi(z'))$, where $\psi : \mathcal{Z} \to \mathbb{R}^p$ is a fixed bounded measurable map. Enclose its image in a fixed compact box $D \subset \mathbb{R}^p$. For every sequence $z_1, \dots, z_m$, the kernel matrix equals that of $\widetilde{k}$ evaluated at $\psi(z_1), \dots, \psi(z_m) \in D$. Thus

$$\Gamma_N(k) \le \sup_{x_1,\dots,x_{NT} \in D} \frac{1}{2} \log\det\Big(I_{NT} + \sigma^{-2}[\widetilde{k}(x_i, x_j)]_{i,j=1}^{NT}\Big). \tag{38}$$

This is a deterministic comparison over all input sequences, so it also covers adaptive trajectories. For fixed D, dimension, kernel parameters, and $\sigma^2 > 0$, Srinivas et al. (2012, Theorem 5) gives

$$\Gamma_N(k) = \begin{cases} \mathcal{O}\big((\log(NT))^{p+1}\big), & \text{squared-exponential,} \\ \mathcal{O}\big((NT)^{p(p+1)/(2\nu+p(p+1))} \log(NT)\big), & \text{Matérn, } \nu > 1. \end{cases} \tag{39}$$

Unlike the Lipschitz feature condition used in Vignola et al. (2026, Appendix A.1), no feature smoothness is needed here. The information-gain comparison uses only the kernel matrices, and Lemma A.1 differentiates a Gaussian-smoothed continuation with respect to its next-state mean, not the feature map. Assumption 3.2 must still hold for the composed kernel. Boundedness of $k(z, z)$ alone does not imply these compact-domain rates.

Information gain in equation 9 permits repeated inputs and updates the scalar posterior after each observation. The following bound accounts for freezing it within an episode.

Lemma A.10 (Episode-frozen posterior variance)**.** *Let Assumption 3.2 hold and model each dynamics coordinate by a scalar GP with kernel k and noise variance $\sigma^2 > 0$, fit on all previous episodes. Let $z_{n,t}$, $n = 1, \dots, N$ and $t = 0, \dots, T-1$, be any realized sequence of inputs, and suppose $k(z, z) \le \sigma_{\max}$. Then*

$$\sum_{n=1}^{N} \sum_{t=0}^{T-1} \|\sigma_{n-1}(z_{n,t})\|_2^2 \le C_{\mathrm{IG}} \Gamma_N(k), \qquad C_{\mathrm{IG}} \triangleq 2 d_\mathcal{S} T(\sigma^2 + \sigma_{\max}), \tag{40}$$

$$\sum_{n=1}^{N} \sum_{t=0}^{T-1} \|\sigma_{n-1}(z_{n,t})\|_2 \le \sqrt{NT\, C_{\mathrm{IG}} \Gamma_N(k)}. \tag{41}$$

Proof. Fix a stage t and coordinate j. Let $\widetilde{\sigma}_{n-1}^{(t)}$ be the scalar posterior standard deviation using only the earlier inputs $z_{1,t}, \dots, z_{n-1,t}$ at that stage. Conditioning on the additional data from the other stages can only reduce posterior variance, so

$$\sum_{n=1}^{N} \log\big(1 + \sigma^{-2} \sigma_{n-1,j}(z_{n,t})^2\big) \le \sum_{n=1}^{N} \log\Big(1 + \sigma^{-2} [\widetilde{\sigma}_{n-1}^{(t)}(z_{n,t})]^2\Big) \le 2\Gamma_N(k).$$

The last inequality is the sequential GP information identity for the N inputs at stage t; $\Gamma_N(k)$ permits up to NT inputs. For $0 \le x^2 \le \sigma_{\max}$,

$$x^2 \le (\sigma^2 + \sigma_{\max}) \log\big(1 + \sigma^{-2} x^2\big).$$

Using $\sigma_{n-1,j}^2 \le \sigma_{\max}$ and summing over stages and coordinates gives equation 40. Cauchy–Schwarz then gives equation 41. □

Lemma A.11 (Multiplicative martingale bound)**.** *Let $(X_n)_{n=1}^N$ be adapted to $(\mathcal{F}_n)_{n=0}^N$, with $X_n \in [0, b]$, and write $m_n = \mathbb{E}[X_n \mid \mathcal{F}_{n-1}]$. Then, for every $\eta \in (0, 1)$, with probability at least $1 - \eta$,*

$$\sum_{n=1}^{N} m_n \le \frac{1}{1 - e^{-1}} \left(\sum_{n=1}^{N} X_n + b \log \frac{1}{\eta} \right). \tag{42}$$

Proof. The case $b = 0$ is immediate. For $b > 0$, put $c = 1 - e^{-1}$. Convexity gives $e^{-x/b} \le 1 - cx/b$ on $[0, b]$, so

$$\mathbb{E}\Big[e^{(cm_n - X_n)/b} \mid \mathcal{F}_{n-1}\Big] \le e^{cm_n/b}(1 - cm_n/b) \le 1.$$

Thus $\mathcal{M}_0 = 1$ and

$$\mathcal{M}_n \triangleq \exp\left(\frac{1}{b}\sum_{i=1}^{n}(cm_i - X_i)\right)$$

form a nonnegative supermartingale. Markov's inequality gives $\mathbb{P}(\mathcal{M}_N \ge 1/\eta) \le \eta$; taking logarithms on the complementary event proves equation 42. □

Corollary A.12 (Expected posterior-uncertainty sums)**.** *Under the conditions of Lemma A.10, suppose σ_{n-1} is $\mathcal{F}_{n-1}$-measurable and $z_{n,0:T-1}$ is $\mathcal{F}_n$-measurable. Define*

$$S_N^\sigma \triangleq \sum_{n=1}^{N}\mathbb{E}\left[\sum_{t=0}^{T-1}\|\sigma_{n-1}(z_{n,t})\|_2 \;\middle|\; \mathcal{F}_{n-1}\right],$$
$$V_N^\sigma \triangleq \sum_{n=1}^{N}\mathbb{E}\left[\sum_{t=0}^{T-1}\|\sigma_{n-1}(z_{n,t})\|_2^2 \;\middle|\; \mathcal{F}_{n-1}\right].$$

For every $\eta \in (0, 1)$, each bound below holds with probability at least $1 - \eta$; both hold with probability at least $1 - 2\eta$:

$$V_N^\sigma \le \frac{1}{1 - e^{-1}}\left[C_{\mathrm{IG}}\Gamma_N(k) + Td_{\mathcal{S}}\sigma_{\max}\log\frac{1}{\eta}\right], \tag{43}$$

$$S_N^\sigma \le \frac{1}{1 - e^{-1}}\left[\sqrt{NT\,C_{\mathrm{IG}}\Gamma_N(k)} + T\sqrt{d_{\mathcal{S}}\sigma_{\max}}\log\frac{1}{\eta}\right]. \tag{44}$$

Proof. Apply Lemma A.11 first to $X_n = \sum_t\|\sigma_{n-1}(z_{n,t})\|_2^2$, which is bounded by $Td_{\mathcal{S}}\sigma_{\max}$, and then to $X_n = \sum_t\|\sigma_{n-1}(z_{n,t})\|_2$, which is bounded by $T\sqrt{d_{\mathcal{S}}\sigma_{\max}}$. Bound their realized sums with equation 40 and equation 41 and union bound. □

Lemma A.13 (Accumulation of weighted posterior uncertainty)**.** *Under the conditions of Lemma A.10, let $w_{t+1,j} : \mathcal{S} \times \mathcal{A} \to [0, \overline{w}]$ be fixed measurable weights, and let the query sequence be generated by policies chosen using the information available before each episode. For the realized query history $\mathcal{H}_N = (z_{n,t})_{n,t}$, define*

$$\mathcal{A}_N(w) \triangleq \sum_{n=1}^{N}\mathbb{E}\left[\sum_{t=0}^{T-1}\sum_{j=1}^{d_{\mathcal{S}}} w_{t+1,j}(z_{n,t})\sigma_{n-1,j}(z_{n,t}) \;\middle|\; \mathcal{F}_{n-1}\right],$$
$$\widehat{\Gamma}_N(w; k, \mathcal{H}_N) \triangleq \frac{1}{2}\sum_{n=1}^{N}\sum_{t=0}^{T-1}\sum_{j=1}^{d_{\mathcal{S}}}\log\big(1 + \sigma^{-2}[w_{t+1,j}(z_{n,t})\sigma_{n-1,j}(z_{n,t})]^2\big).$$

Set $c_0 \triangleq 1 - e^{-1}$ and $C_w \triangleq \sigma^2 + \overline{w}^2\sigma_{\max}$. Then, for every $\eta \in (0, 1)$, with probability at least $1 - 2\eta$, equation 43 holds together with

$$\mathcal{A}_N(w) \le \frac{1}{c_0}\left[\sqrt{2NTd_{\mathcal{S}}C_w\widehat{\Gamma}_N(w; k, \mathcal{H}_N)} + Td_{\mathcal{S}}\overline{w}\sqrt{\sigma_{\max}}\log\frac{1}{\eta}\right]. \tag{45}$$

Writing $\Gamma_N(w; k) \triangleq \sup_{\mathcal{H}_N}\widehat{\Gamma}_N(w; k, \mathcal{H}_N)$, the same event also satisfies

$$\mathcal{A}_N(w) \le \frac{\sqrt{2d_{\mathcal{S}}C_w}}{c_0}\left(\sqrt{NT} + T\log\frac{1}{\eta}\right)\sqrt{\Gamma_N(w; k)},$$
$$V_N^\sigma \le \frac{C_{\mathrm{IG}}}{c_0}\left(1 + \log\frac{1}{\eta}\right)\Gamma_N(k). \tag{46}$$

These bounds depend only on the upper kernel and weight bounds, without a lower bound on the kernel's diagonal.

Proof. Let $\kappa = \sup_z k(z,z)$ and $q_{\max} = \sup_{t,z} \sum_j w_{t+1,j}(z)\sqrt{k(z,z)}$. For a realized episode, let

$$X_n^w \triangleq \sum_{t=0}^{T-1} \sum_{j=1}^{d_{\mathcal{S}}} w_{t+1,j}(z_{n,t})\sigma_{n-1,j}(z_{n,t}).$$

This variable is $\mathcal{F}_n$-measurable, $0 \le X_n^w \le T q_{\max} \le T d_{\mathcal{S}} \overline{w}\sqrt{\sigma_{\max}}$, and its conditional mean is the episode-n summand of $\mathcal{A}_N(w)$. For $0 \le x \le \overline{w}\sqrt{\sigma_{\max}}$,

$$x^2 \le C_w \log(1 + \sigma^{-2}x^2).$$

Cauchy–Schwarz therefore gives the pathwise bound

$$\sum_{n=1}^{N} X_n^w \le \sqrt{2NTd_{\mathcal{S}}C_w \widehat{\Gamma}_N(w;k,\mathcal{H}_N)}.$$

Apply Lemma A.11 to X_n^w with range $Tq_{\max}$ and to $\sum_t \|\sigma_{n-1}(z_{n,t})\|_2^2$ with range $Td_{\mathcal{S}}\kappa$. Together with equation 40, a union bound proves equation 45 and equation 43 on an event of probability at least $1 - 2\eta$.

For the sharper worst-case bounds, one input at stage t of the first episode contributes its prior-variance terms to $\Gamma_N(w;k)$. Cauchy–Schwarz and the same variance-to-log comparison therefore give $q_{\max}^2 \le 2d_{\mathcal{S}}C_w\Gamma_N(w;k)$. Similarly, one input in the ordinary information gain gives $\kappa \le 2(\sigma^2 + \sigma_{\max})\Gamma_N(k)$. Substituting these into the two martingale ranges proves equation 46 on the same event. □

A.5 Accumulated GP prediction error

To apply the simulation bound in Lemma A.9 across episodes, we need to control the accumulated squared prediction error of the GP mean. Calibration and the posterior-variance bounds above give $\widetilde{\mathcal{O}}(\beta_{N-1}^2\Gamma_N(k))$. A direct analysis improves this to $\widetilde{\mathcal{O}}(B^2 + \Gamma_N(k))$. We first combine the online kernel ridge-regression identity of Zhdanov & Kalnishkan (2011, Theorem 1) with a Gaussian martingale argument for sequential predictions (Lemma A.14), then extend the result to posteriors held fixed within each episode and conditional expectations of clipped squared errors (Lemma A.15).

Lemma A.14 (Sequential GP prediction error)**.** *Let $Y_i = f^\star(z_i) + \epsilon_i$, where $\|f^\star\|_k \le B$, $k(z,z) \le \sigma_{\max}$, and $\epsilon_i \sim \mathcal{N}(0,\sigma^2)$ is independent of the data available when z_i is selected. Let $\widetilde{\mu}_{i-1}$ and $\widetilde{\sigma}_{i-1}$ be the GP mean and latent standard deviation in equation 5 using the first $i-1$ observations and set $\gamma_m = \frac{1}{2}\log\det(I + \sigma^{-2}K_m)$. For every $\eta \in (0,1)$, with probability at least $1-\eta$, simultaneously for all m,*

$$\sum_{i=1}^{m} [\widetilde{\mu}_{i-1}(z_i) - f^\star(z_i)]^2 \le (2\sigma^2 + \sigma_{\max})\left[B^2 + 2\gamma_m + 4\log\frac{1}{\eta}\right]. \tag{47}$$

The inputs may be chosen adaptively.

Proof. Write $e_i = \widetilde{\mu}_{i-1}(z_i) - f^\star(z_i)$. Define the conditional Gaussian densities $p_i = \mathcal{N}(f^\star(z_i), \sigma^2)$ and $q_i = \mathcal{N}(\widetilde{\mu}_{i-1}(z_i), \sigma^2 + \widetilde{\sigma}_{i-1}^2(z_i))$. To derive the online ridge identity, set $A_i = K_i + \sigma^2 I$. Its block form is

$$A_i = \begin{pmatrix} A_{i-1} & k_{i-1}(z_i) \\ k_{i-1}(z_i)^\mathsf{T} & k(z_i,z_i) + \sigma^2 \end{pmatrix}.$$

By equation 5, the Schur complement of A_{i-1} is $\sigma^2 + \widetilde{\sigma}_{i-1}^2(z_i)$. The block determinant and inverse formulas give

$$\frac{\det A_i}{\det A_{i-1}} = \sigma^2 + \widetilde{\sigma}_{i-1}^2(z_i),$$

$$Y_{1:i}^\mathsf{T} A_i^{-1} Y_{1:i} - Y_{1:i-1}^\mathsf{T} A_{i-1}^{-1} Y_{1:i-1} = \frac{[Y_i - \widetilde{\mu}_{i-1}(z_i)]^2}{\sigma^2 + \widetilde{\sigma}_{i-1}^2(z_i)}.$$

Use $\det A_0 = 1$ and an empty quadratic form of zero. For $\alpha = A_m^{-1} Y_{1:m}$, the reproducing property and $2ab \le a^2 + b^2$ give

$$\begin{aligned} 2Y_{1:m}^\top \alpha &= 2\left\langle f^\star, \sum_{i=1}^m \alpha_i k(z_i, \cdot)\right\rangle_k + 2\epsilon_{1:m}^\top \alpha \\ &\le B^2 + \alpha^\top K_m \alpha + \sigma^{-2}\|\epsilon_{1:m}\|_2^2 + \sigma^2\|\alpha\|_2^2 \\ &= B^2 + \sigma^{-2}\|\epsilon_{1:m}\|_2^2 + Y_{1:m}^\top \alpha. \end{aligned}$$

Rearranging bounds $Y_{1:m}^\top A_m^{-1} Y_{1:m}$ by $B^2 + \sigma^{-2}\|\epsilon_{1:m}\|_2^2$. Expanding the Gaussian log densities and summing the two recursions therefore gives

$$\begin{aligned} L_m \triangleq \sum_{i=1}^m \log \frac{p_i(Y_i)}{q_i(Y_i)} &= \frac{1}{2}\sum_{i=1}^m \left[\log\left(1 + \frac{\widetilde{\sigma}_{i-1}^2(z_i)}{\sigma^2}\right) + \frac{[Y_i - \widetilde{\mu}_{i-1}(z_i)]^2}{\sigma^2 + \widetilde{\sigma}_{i-1}^2(z_i)} - \frac{\epsilon_i^2}{\sigma^2}\right] \\ &= \gamma_m + \frac{1}{2}\left[Y_{1:m}^\top A_m^{-1} Y_{1:m} - \sigma^{-2}\|\epsilon_{1:m}\|_2^2\right] \\ &\le \gamma_m + \frac{B^2}{2}. \end{aligned} \tag{48}$$

For $a_i = \int \sqrt{p_i(y) q_i(y)}\, dy$, the Gaussian affinity formula gives $-\log a_i \ge e_i^2/[4(2\sigma^2 + \widetilde{\sigma}_{i-1}^2(z_i))]$. The product $\prod_{i=1}^m \frac{\sqrt{q_i(Y_i)/p_i(Y_i)}}{a_i}$ is a nonnegative martingale starting at one. Ville's inequality therefore gives, with probability at least $1-\eta$, simultaneously for all m,

$$\sum_{i=1}^m \frac{e_i^2}{4(2\sigma^2 + \widetilde{\sigma}_{i-1}^2(z_i))} \le \sum_{i=1}^m (-\log a_i) \le \frac{L_m}{2} + \log\frac{1}{\eta}.$$

Substituting equation 48 and $\widetilde{\sigma}_{i-1}^2(z_i) \le \sigma_{\max}$ proves the claim. □

Lemma A.15 (Accumulated GP prediction error)**.** *Under Assumptions 3.1 and 3.2, use the GP posterior equation 5, fitted on all earlier episodes and held fixed within each episode. Define*

$$\ell_n(z) \triangleq \min\left\{\|\mu_{n-1}(z) - f^\star(z)\|_2^2, \sigma^2\right\}. \tag{49}$$

For any policy sequence $(\pi_n)_{n=1}^N$ chosen using the information available before each episode and any $\eta \in (0,1)$, with probability at least $1-\eta$,

$$\sum_{n=1}^N J(\pi_n, f^\star; \ell_n) \le C_{\mathrm{pred}}\left[B^2 + \Gamma_N(k) + \log\frac{3d_{\mathcal{S}}}{\eta}\right]\log\frac{3Nd_{\mathcal{S}}}{\eta}, \tag{50}$$

where C_{pred} depends only on $T, d_{\mathcal{S}}, \sigma$, and $\sigma_{\max}$. Each return on the left is conditional on the data $\mathcal{F}_{n-1}$ available before episode n.

Proof. Within-episode comparison. For episode n, let $Z_n = (z_{n,0}, \dots, z_{n,T-1})$ be its input list and $Z_{<n} = \mathcal{Z}_{1:n-1}$ the list of inputs from earlier episodes, both in chronological order. For two input lists U, V, write $K(U,V) = [k(u,v)]_{u\in U, v\in V}$, so $K(Z_{<n}, Z_{<n}) = K_{n-1}$. Evaluating the posterior from equation 5 at Z_n gives the mean vector for coordinate j and the covariance matrix

$$\begin{aligned} \mu_{n-1,j}(Z_n) &= K(Z_n, Z_{<n})(K_{n-1} + \sigma^2 I)^{-1} y_{1:n-1}^j, \\ C_n &= K(Z_n, Z_n) - K(Z_n, Z_{<n})(K_{n-1} + \sigma^2 I)^{-1} K(Z_{<n}, Z_n). \end{aligned}$$

The rows and columns of $C_n \in \mathbb{R}^{T\times T}$ are indexed by $t = 0, \dots, T-1$; its diagonal is $C_n[t,t] = \sigma_{n-1,j}^2(z_{n,t})$, the same for every coordinate j.

Let $\widetilde{\mu}_{i,j}$ be the posterior mean for coordinate j after the first i transitions, as in Lemma A.14. Define the episode-start error, sequential error, and noise vectors $e_{n,j}, \widetilde{e}_{n,j}, \epsilon_{n,j} \in \mathbb{R}^T$ by their entries:

$$\begin{aligned} (e_{n,j})_t &= \mu_{n-1,j}(z_{n,t}) - f_j^\star(z_{n,t}), \\ (\widetilde{e}_{n,j})_t &= \widetilde{\mu}_{(n-1)T+t,j}(z_{n,t}) - f_j^\star(z_{n,t}), \\ (\epsilon_{n,j})_t &= s_{n,t+1,j} - f_j^\star(z_{n,t}), \qquad t = 0, \dots, T-1. \end{aligned}$$

For the next calculation, fix n, j and write $C = C_n$ and $e, \widetilde{e}, \epsilon$ for these three vectors. Use $< t = \{0, \dots, t-1\}$ to select vector entries or matrix rows and columns. The GP update using the first t observations of the episode gives

$$\widetilde{e}_t - e_t = C[t, <t]\big(C[<t, <t] + \sigma^2 I\big)^{-1}(\epsilon_{<t} - e_{<t}),$$

with a zero update at $t = 0$. Nonnegativity of the updated posterior variance gives

$$C[t, <t](C[<t, <t] + \sigma^2 I)^{-1} C[<t, t] \le C[t, t].$$

Using $(C[<t, <t] + \sigma^2 I)^{-1} \preceq \sigma^{-2} I$, Cauchy–Schwarz and summation over t give

$$|\widetilde{e}_t - e_t|^2 \le \frac{C[t,t]}{\sigma^2}\|\epsilon - e\|_2^2, \qquad \|\widetilde{e} - e\|_2^2 \le \frac{\operatorname{tr} C}{\sigma^2}\|\epsilon - e\|_2^2.$$

Thus

$$\|e\|_2^2 \le 2\|\widetilde{e}\|_2^2 + \frac{4 \operatorname{tr} C}{\sigma^2}\big(\|\epsilon\|_2^2 + \|e\|_2^2\big).$$

If $\operatorname{tr} C \le \sigma^2/8$, moving the last term to the left yields $\|e\|_2^2 \le 4\|\widetilde{e}\|_2^2 + 8\sigma^{-2} \operatorname{tr} C \, \|\epsilon\|_2^2$. If $\operatorname{tr} C > \sigma^2/8$, clipping gives

$$\sum_{t=0}^{T-1} \min\{e_t^2, \sigma^2\} \le T\sigma^2 < 8T \operatorname{tr} C.$$

Both cases therefore imply

$$\sum_{t=0}^{T-1} \min\{e_t^2, \sigma^2\} \le 4\|\widetilde{e}\|_2^2 + 8 \operatorname{tr} C \left(T + \frac{\|\epsilon\|_2^2}{\sigma^2}\right). \tag{51}$$

Accumulation over episodes. The GP posterior covariance depends only on the inputs, kernel, and noise variance, all shared across output coordinates. Thus C_n is the same for every j, and the sum over j counts $d_{\mathcal{S}}$ identical traces:

$$\sum_{n=1}^{N}\sum_{j=1}^{d_{\mathcal{S}}} \operatorname{tr} C_n = \sum_{n=1}^{N}\sum_{t=0}^{T-1} \|\sigma_{n-1}(z_{n,t})\|_2^2.$$

Using $\min\{\sum_j x_j, \sigma^2\} \le \sum_j \min\{x_j, \sigma^2\}$ for $x_j \ge 0$, sum equation 51 over episodes and coordinates to obtain

$$\sum_{n,t} \ell_n(z_{n,t}) \le 4\sum_{n,j} \|\widetilde{e}_{n,j}\|_2^2 + 8\left(T + \max_{n,j} \frac{\|\epsilon_{n,j}\|_2^2}{\sigma^2}\right)\sum_{n,t} \|\sigma_{n-1}(z_{n,t})\|_2^2.$$

Lemma A.14, applied to each coordinate with failure probability $\eta/(3d_{\mathcal{S}})$, bounds $\sum_{n,j} \|\widetilde{e}_{n,j}\|_2^2$ by a constant times $B^2 + \Gamma_N(k) + \log(3d_{\mathcal{S}}/\eta)$, using $\gamma_{NT} \le \Gamma_N(k)$. Lemma A.10 bounds the variance sum by $C_{\mathrm{IG}}\Gamma_N(k)$. A Gaussian norm bound and a union bound give

$$\max_{n,j} \frac{\|\epsilon_{n,j}\|_2^2}{\sigma^2} \le T + 2\sqrt{T \log \frac{3Nd_{\mathcal{S}}}{\eta}} + 2\log \frac{3Nd_{\mathcal{S}}}{\eta}$$

with probability at least $1 - \eta/3$. Substitution bounds the accumulated clipped error $\sum_{n,t} \ell_n(z_{n,t})$ by the right-hand side of equation 50, up to a constant.

Conditional expectations. Set $X_n = \sum_t \ell_n(z_{n,t})$. Then $0 \le X_n \le T\sigma^2$ and $\mathbb{E}[X_n \mid \mathcal{F}_{n-1}] = J(\pi_n, f^\star; \ell_n)$. Lemma A.11, with failure probability $\eta/3$, gives

$$\sum_{n=1}^{N} J(\pi_n, f^\star; \ell_n) \le \frac{1}{1 - e^{-1}}\left[\sum_{n=1}^{N} X_n + T\sigma^2 \log \frac{3}{\eta}\right].$$

The preceding bound on $\sum_n X_n$ and a union bound complete the proof. □

A.6 Common setting for the regret analyses

Let $\mathcal{F}_{n-1}$ contain the data available before episode n, and let $z_{n,t} = (s_{n,t}, a_{n,t})$ be the state–action pair observed when the selected policy is run on $f^\star$. Selected policies and posterior quantities are $\mathcal{F}_{n-1}$-measurable. The process noise and random draws used to sample actions in episode n are independent of $\mathcal{F}_{n-1}$. Fix $\delta \in (0,1)$, use $\beta_{n-1} = \beta_{n-1}(\delta/3)$, and work on the corresponding calibration event unless stated otherwise.

Because the Gaussian noise has full support, take $\mathcal{S} = \mathbb{R}^{d_\mathcal{S}}$. Policies may be time-dependent Markov policies.

Write $H_t \triangleq R_{\max}(T-t)$. For the posterior standard-deviation vector, write

$$u_{n-1}(z) \triangleq \|\sigma_{n-1}(z)\|_2, \qquad 0 \le u_{n-1}(z) \le \sqrt{d_\mathcal{S}}\sigma_{\max}. \tag{52}$$

For asymptotic rate statements, the kernel, horizon T, and other problem parameters are fixed. For a nonzero kernel, choose z_0 with $k(z_0, z_0) > 0$. Then $\Gamma_N(k) \ge \frac{1}{2}\log(1 + \sigma^{-2}k(z_0, z_0)) > 0$, so $1 + \Gamma_N(k) = \mathcal{O}(\Gamma_N(k))$, with a constant depending on the fixed kernel and noise variance. This absorbs additive concentration terms in the rates below. If $k \equiv 0$, realizability makes the dynamics known and the optimistic planners have zero regret.

B Optimistic Synthesis and Hallucinated Control

Optimistic synthesis chooses a policy together with the most favorable plausible dynamics. Using the common setting of Appendix A.6, we express this choice as hallucinated control and compare its one-step optimistic bonus with the value-directed bonus of this paper and the undirected exploration bonus of SOMBRL (Sukhija et al., 2025b).

B.1 Exact optimistic synthesis

At episode n, define the coordinatewise confidence radii

$$\rho_{n,j}(z) \triangleq \beta_{n-1}\sigma_{n-1,j}(z). \tag{53}$$

The corresponding class of plausible dynamics maps is

$$\mathcal{G}_n \triangleq \{g : \mathcal{S} \times \mathcal{A} \to \mathcal{S} \text{ measurable} : |g_j(z) - \mu_{n-1,j}(z)| \le \rho_{n,j}(z) \text{ for every } (z,j)\}. \tag{54}$$

Exact optimistic synthesis selects

$$(\pi_n^{\mathrm{ios}}, g_n) \in \arg\max_{\pi\in\Pi,\, g\in\mathcal{G}_n} J(\pi, g; r). \tag{55}$$

On calibration, $f^\star \in \mathcal{G}_n$, so $J(\pi_n^{\mathrm{ios}}, g_n; r) \ge J(\pi^\star, f^\star; r)$.

Equivalently, parameterize the plausible dynamics by

$$g_{n,\eta}(z) \triangleq \mu_{n-1}(z) + \rho_n(z) \odot \eta(z), \qquad \eta(z) \in [-1,1]^{d_\mathcal{S}}. \tag{56}$$

Every measurable η gives $g_{n,\eta} \in \mathcal{G}_n$. Conversely, any $g \in \mathcal{G}_n$ is obtained by setting $\eta_j(z) = (g_j(z) - \mu_{n-1,j}(z))/\rho_{n,j}(z)$ when $\rho_{n,j}(z) > 0$, and $\eta_j(z) = 0$ otherwise. Thus equation 55 optimizes jointly over (π, η). This is exactly the H-UCRL objective of Curi et al. (2020, Eq. (7)), with η providing $d_\mathcal{S}$ hallucinated control coordinates.

Connection to LC^3. The same optimism argument holds when synthesis is restricted to any $\mathcal{F}_{n-1}$-measurable model class $\mathcal{M}_n \subseteq \mathcal{G}_n$ containing $f^\star$ on the calibration event. The LC^3 algorithm of Kakade et al. (2020, Algorithm 1 and Eq. (3.2)) jointly optimizes over policies and such a class, defined by a confidence ellipsoid for the dynamics coefficients.

B.2 Optimistic exploration bonuses

Fix a bounded measurable continuation $F : \mathbb{R}^{d_\mathcal{S}} \to [0, H]$ and a state–action pair z. With $x = \mu_{n-1}(z)$ and $\rho = \rho_n(z)$, optimizing H-UCRL's hallucinated control at this step gives the continuation

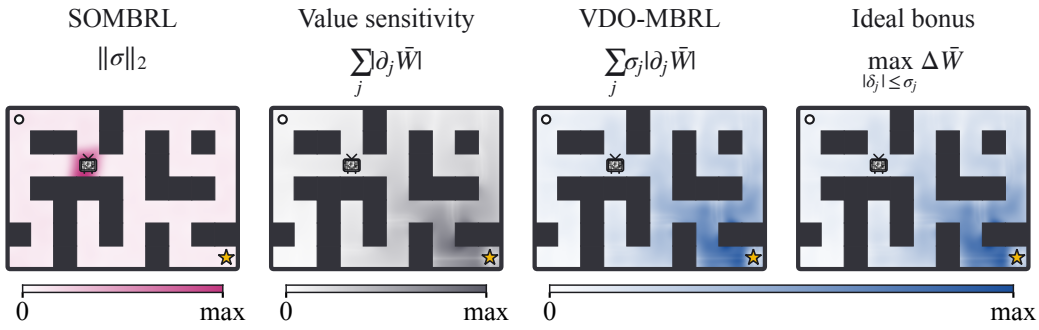


Figure 4: Heatmaps of the relevant terms at each maze position averaged over actions under the reference policy. The circle and star mark the start and goal positions respectively.

backup $(P_{\mu_{n-1}} F)(z) + \Omega_\rho^+[F](x)$. The confidence width $\Omega_\rho^+[F](x)$ from Corollary A.4 is therefore an exact optimistic exploration bonus, added to the reward under the posterior-mean dynamics. The Taylor and SOMBRL bonuses approximate this same bonus from above.

The Taylor approximation of this bonus satisfies

$$0 \le L_\rho[F](x) + Q_\rho(H) - \Omega_\rho^+[F](x) \le 2Q_\rho(H) = \frac{c_H H}{\sigma^2} \|\rho\|_2^2 \tag{57}$$

by equation 28, so its gap is second order in the confidence radius. Its linear term $L_\rho[F](x)$ uses the local gradient $\nabla \overline{F}(x)$. SOMBRL (Sukhija et al., 2025b) instead uses the undirected bound equation 31, which gives $\Omega_\rho^+[F](x) \le (H/\sigma)\beta_{n-1} u_{n-1}(z)$. Taking $H = R_{\max} T$ recovers its exploration bonus, which uses only the magnitude of model uncertainty.

Exact synthesis chooses one dynamics map g_n, so $g_n(s, a)$ must be the same at every stage. The maximization defining the one-step optimistic bonus $\Omega_\rho^+[F](x)$ associated with H-UCRL can favor different next-state means for the same (s, a) at different stages, because the continuation F changes.[1]

Comparison of optimistic bonuses. A comparison of the bonuses can be need in Figure 4 for the noisy TV example. SOMBRL assigns a large bonus near the TV because uncertainty about the display dynamics increases model uncertainty regardless of its relevance to the task. Since the learned value is independent of the display, first-order value weighting removes this task-irrelevant contribution and instead encourages exploration along the approach to the goal where prediction errors have a greater effect on future return. The similar sensitivity and VDO densities reflect approximately uniform uncertainty in the task dynamics in this example. The two rightmost panels compare $L_\rho[W]$ with $\Omega_\rho^+[W]$ using the same value function and their close agreement supports that gradient weighting approximates the largest increase in future value across plausible next-state predictions.

C Self-Consistent Taylor Optimism

We use the common setting of Appendix A.6 and replace the exact one-step confidence support from Appendix B by a Taylor upper bound. The resulting bonus is evaluated using the continuation that it defines. We prove a regret bound for any policy sequence with sufficiently high scores, and verify this condition for ideal optimistic synthesis, global score maximization, and a backward construction.

[1] For stagewise tabular confidence sets, this model/value optimism connection is formalized by Neu and Pike-Burke (2020), *A Unifying View of Optimism in Episodic Reinforcement Learning*.

C.1 Taylor elimination of hallucinated control

Fix a measurable continuation $F : \mathcal{S} \to [0, H]$ and a state–action pair z. Linearizing the perturbation $\rho_n(z) \odot \eta$ in equation 56 at the posterior mean gives

$$\sup_{\eta \in [-1,1]^{d_{\mathcal{S}}}} \nabla \overline{F}(\mu_{n-1}(z))^\top (\rho_n(z) \odot \eta) = \sum_{j=1}^{d_{\mathcal{S}}} \rho_{n,j}(z) |\partial_j \overline{F}(\mu_{n-1}(z))|, \tag{58}$$

attained coordinatewise by $\eta_j = \operatorname{sign}(\partial_j \overline{F}(\mu_{n-1}(z)))$. Corollary A.4 bounds the discarded nonlinear part by $c_H H \|\rho_n(z)\|_2^2/(2\sigma^2)$, leading to the VDO bonus.

C.2 Policywise Taylor certificate and synthesis criterion

For episode n and candidate policy $\pi \in \Pi$, define the following backward recursion, where $\overline{\mathsf{U}}^\pi_{n,t}(y) = \mathbb{E}_\epsilon[\mathsf{U}^\pi_{n,t}(y + \epsilon)]$ is the Gaussian smoothing from equation 21:

$$\mathsf{U}^\pi_{n,T} \equiv 0,$$

$$\mathsf{G}^\pi_{n,t+1,j}(z) \triangleq \left| \partial_j \overline{\mathsf{U}}^\pi_{n,t+1}(\mu_{n-1}(z)) \right|, \tag{59}$$

$$\mathsf{b}^\pi_{n,t}(z) \triangleq \beta_{n-1} \sum_{j=1}^{d_{\mathcal{S}}} \mathsf{G}^\pi_{n,t+1,j}(z) \sigma_{n-1,j}(z) + \frac{c_H H_{t+1}}{2\sigma^2} \beta_{n-1}^2 u_{n-1}(z)^2, \tag{60}$$

$$\mathsf{U}^\pi_{n,t}(s) \triangleq \min\left\{ H_t, \mathbb{E}_{a \sim \pi_t(\cdot|s)} \left[r(s,a) + (P_{\mu_{n-1}} \mathsf{U}^\pi_{n,t+1})(s,a) + \mathsf{b}^\pi_{n,t}(s,a) \right] \right\}. \tag{61}$$

The recursion gives $0 \le \mathsf{U}^\pi_{n,t} \le H_t$ and uses the same candidate policy π throughout.

Define the policy score by

$$\mathsf{S}_n(\pi) \triangleq \mathbb{E}_{s_0 \sim \rho_0}[\mathsf{U}^\pi_{n,0}(s_0)]. \tag{62}$$

Theorem C.3 will give a regret bound for any sequence $\nu_1, \dots, \nu_N$, with ν_n chosen using the data available before episode n, that satisfies, on the calibration event,

$$\mathsf{S}_n(\nu_n) \ge J(\pi^\star, f^\star; r), \qquad n = 1, \dots, N. \tag{63}$$

We consider three selection rules.

Ideal optimistic synthesis. Use the pair $(\pi^{\mathrm{ios}}_n, g_n)$ from equation 55, which jointly maximizes $J(\pi, g; r)$ over $\pi \in \Pi$ and $g \in \mathcal{G}_n$. More generally, the maximization may use a class $\mathcal{M}_n \subseteq \mathcal{G}_n$ containing $f^\star$ on calibration. This covers the exact LC^3 and H-UCRL objectives identified in Appendix B.

Self-consistent score maximization. Select any measurable maximizer

$$\pi^{\mathrm{sc}}_n \in \arg\max_{\pi \in \Pi} \mathsf{S}_n(\pi). \tag{64}$$

Backward construction. Write $\mathsf{b}_{n,t}[F]$ for equation 60 with continuation F in place of $\mathsf{U}^\pi_{n,t+1}$. Starting from $\mathsf{U}^{\mathrm{bk}}_{n,T} = 0$, form

$$\mathsf{U}^{\mathrm{bk}}_{n,t}(s) = \min\left\{ H_t, \sup_{a \in \mathcal{A}} \left[r(s,a) + (P_{\mu_{n-1}} \mathsf{U}^{\mathrm{bk}}_{n,t+1})(s,a) + \mathsf{b}_{n,t}[\mathsf{U}^{\mathrm{bk}}_{n,t+1}](s,a) \right] \right\}. \tag{65}$$

Suppose the maximizing actions define a measurable policy $\pi^{\mathrm{bk}}_n \in \Pi$.

Verification of the score condition. The common argument is that the Taylor score bounds the return of the same policy under every plausible model. For any $\pi \in \Pi$ and $g \in \mathcal{G}_n$, the Taylor certificate equation 25 gives

$$\left| (P_{\mu_{n-1}} \mathsf{U}^\pi_{n,t+1})(z) - (P_g \mathsf{U}^\pi_{n,t+1})(z) \right| \le \mathsf{b}^\pi_{n,t}(z). \tag{66}$$

Both entries in the minimum defining equation 61 are therefore at least $\mathbb{E}_{a\sim\pi_t}[r + P_g \mathsf{U}^{\pi}_{n,t+1}]$: for the cap, this follows from $r \le R_{\max}$ and $\mathsf{U}^{\pi}_{n,t+1} \le H_{t+1}$; for the other entry, it follows from equation 66. Backward induction from the common terminal value zero yields

$$\mathsf{U}^{\pi}_{n,t}(s) \ge V^{\pi,g}_{r,t}(s), \qquad g \in \mathcal{G}_n. \tag{67}$$

On calibration, $f^\star \in \mathcal{M}_n \subseteq \mathcal{G}_n$, taking $\mathcal{M}_n = \mathcal{G}_n$ for the full confidence class. The envelope equation 67 applies to the selected g_n, and optimality over $\Pi \times \mathcal{M}_n$ allows comparison with $(\pi^\star, f^\star)$. Hence

$$\mathsf{S}_n(\pi^{\mathrm{ios}}_n) \ge J(\pi^{\mathrm{ios}}_n, g_n; r) \ge J(\pi^\star, f^\star; r). \tag{68}$$

For LC^3 and H-UCRL, the Taylor score is therefore an analysis device: their original optimistic planning objectives already ensure the required score.

For self-consistent score maximization, apply equation 67 to $\pi^\star$ and $f^\star$ to obtain, on calibration,

$$\mathsf{S}_n(\pi^{\mathrm{sc}}_n) \ge \mathsf{S}_n(\pi^\star) \ge J(\pi^\star, f^\star; r). \tag{69}$$

For the backward construction, the recursion equation 65 gives $\mathsf{U}^{\mathrm{bk}}_{n,t} = \mathsf{U}^{\pi^{\mathrm{bk}}_n}_{n,t}$. On calibration, each mean-model backup plus bonus is at least $r + P_{f^\star}\mathsf{U}^{\mathrm{bk}}_{n,t+1}$. For every $\pi \in \Pi$, backward induction therefore gives $\mathsf{U}^{\mathrm{bk}}_{n,t} \ge V^{\pi,f^\star}_{r,t}$: the action supremum dominates the expectation under π, and clipping preserves the inequality because $V^{\pi,f^\star}_{r,t} \le H_t$. In particular,

$$\mathsf{S}_n(\pi^{\mathrm{bk}}_n) \ge J(\pi^\star, f^\star; r). \tag{70}$$

This policy need not maximize S_n: increasing a continuation pointwise can reduce its gradient bonus, so the Taylor backup need not preserve the ordering of continuation values.

We next bound the true-model residual of the Taylor recursion, which will control the regret of any policy satisfying equation 63.

Lemma C.1 (True-model residual control)**.** *For a fixed or $\mathcal{F}_{n-1}$-measurable policy ν, define*

$$e^{\nu}_{n,t}(s) \triangleq \mathsf{U}^{\nu}_{n,t}(s) - \mathbb{E}_{a\sim\nu_t(\cdot|s)}\left[r(s,a) + (P_{f^\star}\mathsf{U}^{\nu}_{n,t+1})(s,a)\right].$$

On the calibration event, for every t and s,

$$0 \le e^{\nu}_{n,t}(s) \le 2\mathbb{E}_{a\sim\nu_t(\cdot|s)}[\mathsf{b}^{\nu}_{n,t}(s,a)]. \tag{71}$$

Proof. With $g = f^\star$, both entries in the minimum equation 61 are at least $\mathbb{E}_{a\sim\nu_t}[r + P_{f^\star}\mathsf{U}^{\nu}_{n,t+1}]$, so $e^{\nu}_{n,t} \ge 0$. Dropping the cap and using equation 66 gives

$$\begin{aligned} e^{\nu}_{n,t}(s) &\le \mathbb{E}_{a\sim\nu_t(\cdot|s)}\left[(P_{\mu_{n-1}} - P_{f^\star})\mathsf{U}^{\nu}_{n,t+1}(s,a) + \mathsf{b}^{\nu}_{n,t}(s,a)\right] \\ &\le 2\mathbb{E}_{a\sim\nu_t(\cdot|s)}[\mathsf{b}^{\nu}_{n,t}(s,a)]. \end{aligned}$$

□

C.3 Optimal-value relevance

With the transition operator from equation 21, define the unrestricted optimal value under the true dynamics by

$$V^\star_T \equiv 0, \qquad V^\star_t(s) \triangleq \sup_{a\in\mathcal{A}} \left\{r(s,a) + (P_{f^\star}V^\star_{t+1})(s,a)\right\}. \tag{72}$$

Assume that each $V^\star_t$ is measurable. For the remainder of Appendix C, assume that Π attains the Bellman-optimal return from the initial-state distribution:

$$J(\pi^\star, f^\star; r) = \mathbb{E}_{s_0\sim\rho_0}[V^\star_0(s_0)]. \tag{73}$$

Lemma C.2 (High-score certificate)**.** *Let ν be fixed or $\mathcal{F}_{n-1}$-measurable and suppose*

$$\mathsf{S}_n(\nu) \geq J(\pi^\star, f^\star; r).$$

Define

$$B_n(\nu) \triangleq \mathbb{E}_{\nu,f^\star}\left[\sum_{t=0}^{T-1} \mathsf{b}_{n,t}^{\nu}(z_t) \;\middle|\; \mathcal{F}_{n-1}\right].$$

On the calibration event,

$$J(\pi^\star, f^\star; r) - J(\nu, f^\star; r) \leq 2B_n(\nu), \tag{74}$$

$$\mathbb{E}_{\nu,f^\star}\left[\sum_{t=0}^{T-1} (P_{f^\star}|\mathsf{U}_{n,t+1}^{\nu} - V_{t+1}^{\star}|)(z_t) \;\middle|\; \mathcal{F}_{n-1}\right] \leq 4TB_n(\nu). \tag{75}$$

Proof. We drop the notation for conditioning on $\mathcal{F}_{n-1}$ throughout the proof. Telescoping the residuals along the true trajectory gives

$$\mathsf{U}_{n,t}^{\nu}(s) - V_{r,t}^{\nu,f^\star}(s) = \mathbb{E}_{\nu,f^\star}\left[\sum_{\ell=t}^{T-1} e_{n,\ell}^{\nu}(s_\ell) \;\middle|\; s_t = s\right] \geq 0. \tag{76}$$

At $t = 0$, the high-score condition and equation 71 yield

$$\begin{aligned} J(\pi^\star, f^\star; r) - J(\nu, f^\star; r) &\leq \mathsf{S}_n(\nu) - J(\nu, f^\star; r) \\ &= \mathbb{E}_{\nu,f^\star}\left[\sum_t e_{n,t}^{\nu}(s_t)\right] \leq 2B_n(\nu). \end{aligned}$$

Both $\mathsf{U}_{n,t}^{\nu}$ and $V_t^\star$ dominate $V_{r,t}^{\nu,f^\star}$, so

$$|\mathsf{U}_{n,t}^{\nu} - V_t^\star| \leq (\mathsf{U}_{n,t}^{\nu} - V_{r,t}^{\nu,f^\star}) + (V_t^\star - V_{r,t}^{\nu,f^\star}).$$

The first difference is the future residual sum in equation 76. For the second, write $d_\ell(z) \triangleq V_\ell^\star(s) - r(z) - (P_{f^\star}V_{\ell+1}^\star)(z)$. These Bellman gaps are nonnegative by equation 72, and telescoping expresses $V_t^\star - V_{r,t}^{\nu,f^\star}$ as their expected future sum. Over the whole episode, equation 73 gives

$$\begin{aligned} \mathbb{E}_{\nu,f^\star}\left[\sum_{\ell=0}^{T-1} d_\ell(z_\ell)\right] &= \mathbb{E}_{s_0\sim\rho_0}[V_0^\star(s_0)] - J(\nu, f^\star; r) \\ &= J(\pi^\star, f^\star; r) - J(\nu, f^\star; r) \leq 2B_n(\nu). \end{aligned}$$

Applying the tower rule to the two future sums therefore yields

$$\begin{aligned} &\mathbb{E}_{\nu,f^\star}\left[\sum_{t=0}^{T-1} (P_{f^\star}|\mathsf{U}_{n,t+1}^{\nu} - V_{t+1}^{\star}|)(z_t)\right] \\ &\leq \mathbb{E}_{\nu,f^\star}\left[\sum_{t=0}^{T-1}\sum_{\ell=t+1}^{T-1} \big(e_{n,\ell}^{\nu}(s_\ell) + d_\ell(z_\ell)\big)\right] \\ &= \mathbb{E}_{\nu,f^\star}\left[\sum_{\ell=1}^{T-1} \ell\big(e_{n,\ell}^{\nu}(s_\ell) + d_\ell(z_\ell)\big)\right] \\ &\leq T\,\mathbb{E}_{\nu,f^\star}\left[\sum_{\ell=0}^{T-1} \big(e_{n,\ell}^{\nu}(s_\ell) + d_\ell(z_\ell)\big)\right] \leq 4TB_n(\nu). \end{aligned}$$

The factor ℓ counts the earlier stages $t < \ell$. The last inequality uses nonnegativity, $\ell \leq T$, and the bound $2B_n(\nu)$ on each expected total sum. This proves equation 75. □

Let $\overline{V}^\star_{t+1}$ be the Gaussian smoothing of the true optimal value and define

$$w^\star_{t+1,j}(z) \triangleq \left|\partial_j \overline{V}^\star_{t+1}(f^\star(z))\right|, \tag{77}$$

$$0 \le w^\star_{t+1,j}(z) \le w_{\max} \triangleq \frac{c_g R_{\max} T}{\sigma}. \tag{78}$$

These weights measure the task relevance of uncertainty under the true optimal value. For a realized history $\mathcal{H}_N = (z_{n,t})_{n,t}$, define

$$\widehat{\Gamma}^\star_N(k;\mathcal{H}_N) \triangleq \frac{1}{2}\sum_{n=1}^{N}\sum_{t=0}^{T-1}\sum_{j=1}^{d_{\mathcal{S}}} \log\!\left(1+\sigma^{-2}[w^\star_{t+1,j}(z_{n,t})\sigma_{n-1,j}(z_{n,t})]^2\right), \tag{79}$$

$$\Gamma^\star_N(k) \triangleq \sup_{\mathcal{H}_N} \widehat{\Gamma}^\star_N(k;\mathcal{H}_N). \tag{80}$$

The posterior is conditioned on all observations from earlier episodes, and the realized gain uses the deployed policies' query locations. Moreover, equation 78, $\log(1+x) \le x$, and Lemma A.10 give

$$\widehat{\Gamma}^\star_N(k;\mathcal{H}_N) \le \frac{w^2_{\max} C_{\mathrm{IG}}}{2\sigma^2}\Gamma_N(k). \tag{81}$$

The factor depends on the problem because the relevance weights need not be at most one.

To compare these weights with the bonus weights, apply Lemma A.2 to the continuations $\mathsf{U}^\pi_{n,t+1}$ and $V^\star_{t+1}$ at $\mu_{n-1}(z)$ and $f^\star(z)$. For every π, n, t, j, and z, this gives

$$\begin{aligned}\mathsf{G}^\pi_{n,t+1,j}(z) \le w^\star_{t+1,j}(z) &+ \frac{c_H H_{t+1}}{\sigma^2}\|\mu_{n-1}(z) - f^\star(z)\|_2 \\ &+ \frac{1}{\sigma}\sqrt{H_{t+1}(P_{f^\star}|\mathsf{U}^\pi_{n,t+1} - V^\star_{t+1}|)(z)}.\end{aligned} \tag{82}$$

C.4 Regret bound

Theorem C.3 (Regret from high Taylor scores)**.** *Assume Assumptions 3.1, 3.2, and 3.4, the standing conditions of this section, and the GP conditions of Lemma A.13. Let $\beta_n \ge 1$ be the nondecreasing confidence multipliers from Assumption 3.4. Let ν_n be chosen using the data available before episode n, and suppose $\nu_1, \dots, \nu_N$ satisfy equation 63 on the calibration event. Let $\mathcal{H}^\nu_N = (z_{n,t})_{n,t}$ be its realized history and define*

$$R_N(\nu) \triangleq \sum_{n=1}^{N}[J(\pi^\star, f^\star; r) - J(\nu_n, f^\star; r)].$$

Then, with probability at least $1-\delta$,

$$R_N(\nu) = \widetilde{\mathcal{O}}\left(\beta_{N-1}\sqrt{N\,\widehat{\Gamma}^\star_N(k;\mathcal{H}^\nu_N)} + \beta^2_{N-1}\Gamma_N(k)\right), \tag{83}$$

$$R_N(\nu) = \widetilde{\mathcal{O}}\left(\beta_{N-1}\sqrt{N\,\Gamma^\star_N(k)} + \beta^2_{N-1}\Gamma_N(k)\right). \tag{84}$$

Here $\widetilde{\mathcal{O}}$ suppresses logarithmic factors in N and $1/\delta$ and constants depending only on the fixed problem parameters, including the horizon T.

Proof. We drop the notation for conditioning episode-n expectations on $\mathcal{F}_{n-1}$ throughout the proof. Write $\mathsf{U}_{n,t} = \mathsf{U}^{\nu_n}_{n,t}$ and $\beta = \beta_{N-1}$. Define

$$A^\star \triangleq \sum_{n=1}^{N}\mathbb{E}_{\nu_n, f^\star}\left[\sum_{t=0}^{T-1}\sum_{j=1}^{d_{\mathcal{S}}} w^\star_{t+1,j}(z_{n,t})\sigma_{n-1,j}(z_{n,t})\right],$$

$$B \triangleq \sum_{n=1}^{N} B_n(\nu_n),$$

$$V \triangleq \sum_{n=1}^{N} J(\nu_n, f^\star; u^2_{n-1}).$$

Let C denote a positive constant depending only on the fixed problem parameters; its value may change from line to line.

Lemma C.2 gives

$$R_N(\nu) \le 2B, \tag{85}$$

$$\sum_{n=1}^{N} \mathbb{E}_{\nu_n, f^\star}\left[\sum_{t=0}^{T-1} (P_{f^\star}|\mathsf{U}_{n,t+1} - V^\star_{t+1}|)(z_{n,t})\right] \le 4TB. \tag{86}$$

Multiply equation 82 by $\beta_{n-1}\sigma_{n-1,j}(z)$ and sum over coordinates, stages, and episodes. The bonus definition, calibration, monotonicity of β_n, and Cauchy–Schwarz give

$$\begin{aligned} B \le \beta A^\star + \frac{c_H R_{\max} T(1+2\sqrt{d_{\mathcal{S}}})}{2\sigma^2}\beta^2 V \\ + \frac{2T\sqrt{d_{\mathcal{S}} R_{\max}}}{\sigma}\beta\sqrt{VB}, \end{aligned} \tag{87}$$

where we used $\|\sigma_{n-1}(z)\|_1 \le \sqrt{d_{\mathcal{S}}}u_{n-1}(z)$ and equation 86. Young's inequality now gives

$$B \le \beta A^\star + C\beta^2 V + \frac{1}{2}B.$$

Rearranging and using equation 85 gives

$$R_N(\nu) \le 2B \le C(\beta A^\star + \beta^2 V). \tag{88}$$

The curvature bonus and the comparison with $V^\star$ both contribute to $\beta^2 V$.

Apply Lemma A.13 with $w = w^\star$, $\overline{w} = w_{\max}$, $\mathcal{H}_N = \mathcal{H}_N^\nu$, and $\eta = \delta/3$. With probability at least $1 - 2\delta/3$,

$$\begin{aligned} A^\star &\le \frac{1}{c_0}\left[\sqrt{2NTd_{\mathcal{S}}C_{w^\star}\widehat{\Gamma}^\star_N(k;\mathcal{H}_N^\nu)} + Td_{\mathcal{S}}w_{\max}\sqrt{\sigma_{\max}}\log\frac{3}{\delta}\right], \\ V &\le \frac{1}{c_0}\left[C_{\mathrm{IG}}\Gamma_N(k) + Td_{\mathcal{S}}\sigma_{\max}\log\frac{3}{\delta}\right], \end{aligned}$$

where $c_0 = 1 - e^{-1}$ and $C_{w^\star} = \sigma^2 + w_{\max}^2\sigma_{\max}$. Intersecting these events with calibration and using equation 88 gives

$$R_N(\nu) = \widetilde{\mathcal{O}}\left(\beta\sqrt{N\,\widehat{\Gamma}^\star_N(k;\mathcal{H}_N^\nu)} + \beta\log\frac{3}{\delta} + \beta^2\left[\Gamma_N(k) + \log\frac{3}{\delta}\right]\right).$$

Since $\beta \ge 1$, the fixed-kernel observation in Appendix A.6 absorbs the logarithmic terms into $\widetilde{\mathcal{O}}(\beta^2\Gamma_N(k))$, proving equation 83. For equation 84, substitute the worst-case accumulation bounds equation 46 directly into equation 88. Their constants depend only on the upper kernel and weight bounds and the fixed problem parameters. □

The high-score verifications in equation 68, equation 69, and equation 70 apply the theorem to ideal optimistic synthesis, self-consistent score maximization, and the backward construction, respectively, under their stated selection assumptions. Each rule therefore satisfies equation 83 and equation 84 with probability at least $1 - \delta$. In particular, these bounds apply to exact LC^3 (Kakade et al., 2020) and H-UCRL (Curi et al., 2020).

For the confidence sequence of Chowdhury & Gopalan (2017, Theorem 2), $\beta^2_{N-1} = \widetilde{\mathcal{O}}(\Gamma_N(k))$, so equation 84 gives $R_N = \widetilde{\mathcal{O}}(\sqrt{N\,\Gamma_N(k)\Gamma^\star_N(k)} + \Gamma_N(k)^2)$. The theorem keeps β_{N-1} explicit so that any tighter calibrated confidence sequence improves the corresponding regret bound.

D Task-Only Value-Directed Exploration

This section analyzes an analogue of VDO's task-only value geometry, using the common setting of Appendix A.6 and the relevance quantities from Appendix C. The value-directed term uses the candidate policy's task-only continuation under the posterior-mean dynamics. Future intrinsic reward does not enter that continuation, matching the separation between the extrinsic critic's value gradients and the augmented critic's actor objective.

D.1 TASK-ONLY POLICY CRITERION

For episode n and candidate policy π, define

$$W^{\pi}_{n,t}(s) \triangleq V^{\pi,\mu_{n-1}}_{r,t}(s), \qquad \overline{W}^{\pi}_{n,t}(y) \triangleq \mathbb{E}_{\epsilon}[W^{\pi}_{n,t}(y+\epsilon)], \tag{89}$$

$$G^{\pi}_{n,t+1,j}(z) \triangleq \left|\partial_j \overline{W}^{\pi}_{n,t+1}(\mu_{n-1}(z))\right|. \tag{90}$$

Since $0 \le W^{\pi}_{n,t} \le H_t$, the calibrated Taylor certificate equation 25 gives the one-step continuation-error bound

$$q^{\pi}_{n,t}(z) \triangleq \min\left\{ H_{t+1}, \beta_{n-1} \sum_{j=1}^{d_S} G^{\pi}_{n,t+1,j}(z)\sigma_{n-1,j}(z) + \frac{c_H H_{t+1}}{2\sigma^2}\beta^2_{n-1}u_{n-1}(z)^2 \right\}. \tag{91}$$

To transfer the accumulated one-step bound between the true and mean models, use the constant $C^{\mathrm{task}}_{\mathrm{B}}$ defined in equation 15 and the capped correction

$$v_n(z) \triangleq \min\left\{\beta^2_{n-1}u_{n-1}(z)^2, \sigma^2\right\}. \tag{92}$$

The cap reflects the bounded change in continuation value even when the mean displacement exceeds the noise scale. The full planning bonus is

$$b^{\pi}_{n,t}(z) \triangleq \frac{3}{2}q^{\pi}_{n,t}(z) + \frac{3}{2}C^{\mathrm{task}}_{\mathrm{B}}v_n(z). \tag{93}$$

The certified task-only oracle selects

$$\pi_n \in \arg\max_{\pi\in\Pi} J(\pi, \mu_{n-1}; r + b^{\pi}_n). \tag{94}$$

To compare this score with the true task return, we bound the change in both task returns and accumulated one-step bounds when replacing the mean model by $f^\star$.

Lemma D.1 (Task-only certificate and one-episode regret)**.** *On the calibration event, every fixed or $\mathcal{F}_{n-1}$-measurable policy π, start time τ, and state s satisfy*

$$\left|V^{\pi,f^\star}_{r,\tau}(s) - V^{\pi,\mu_{n-1}}_{r,\tau}(s)\right| \le \mathbb{E}_{\pi,f^\star}\left[\sum_{t=\tau}^{T-1} q^{\pi}_{n,t}(z_t) \,\middle|\, s_\tau = s\right]. \tag{95}$$

In particular,

$$|J(\pi, f^\star; r) - J(\pi, \mu_{n-1}; r)| \le J(\pi, f^\star; q^{\pi}_n), \tag{96}$$

$$J(\pi, f^\star; q^{\pi}_n) \le \frac{3}{2}\left[J(\pi, \mu_{n-1}; q^{\pi}_n) + C^{\mathrm{task}}_{\mathrm{B}}J(\pi, \mu_{n-1}; v_n)\right], \tag{97}$$

$$J(\pi, \mu_{n-1}; q^{\pi}_n) \le 2J(\pi, f^\star; q^{\pi}_n) + 3C^{\mathrm{task}}_{\mathrm{B}}J(\pi, \mu_{n-1}; v_n). \tag{98}$$

Consequently, the selected policy obeys

$$J(\pi^\star, f^\star; r) - J(\pi_n, f^\star; r) \le 4J(\pi_n, f^\star; q^{\pi_n}_n) + 6C^{\mathrm{task}}_{\mathrm{B}}J(\pi_n, \mu_{n-1}; v_n). \tag{99}$$

Proof. The one-step difference satisfies

$$\left|(P_{f^\star}W^{\pi}_{n,t+1})(z) - (P_{\mu_{n-1}}W^{\pi}_{n,t+1})(z)\right| \le q^{\pi}_{n,t}(z)$$

by calibration and equation 25. Corollary A.6, with $f = f^\star$ and $g = \mu_{n-1}$, proves equation 95. Averaging its $\tau = 0$ case over ρ_0 proves equation 96.

To transfer the accumulated one-step bound, apply Lemma A.9 with $(f, g, h) = (f^\star, \mu_{n-1}, \mu_{n-1})$ and the time-indexed reward q^{π}_n. This reward is bounded by $R_{\max}T$. Calibration and the cap give

$$\min\left\{\|f^\star(z) - \mu_{n-1}(z)\|^2_2, \sigma^2\right\} \le v_n(z).$$

The simulation bound therefore gives

$$|J(\pi, f^\star; q^{\pi}_n) - J(\pi, \mu_{n-1}; q^{\pi}_n)| \le C^{\mathrm{task}}_{\mathrm{B}}J(\pi, \mu_{n-1}; v_n) + \sqrt{C^{\mathrm{task}}_{\mathrm{B}}J(\pi, \mu_{n-1}; v_n)J(\pi, \mu_{n-1}; q^{\pi}_n)}.$$

Applying $\sqrt{xy} \le (x+y)/2$ in the two directions proves equation 97 and equation 98.

By equation 96 and equation 97, the planning score upper-bounds every candidate's true task return. Optimality of π_n and equation 96 then give

$$J(\pi^\star, f^\star; r) - J(\pi_n, f^\star; r) \le J(\pi_n, f^\star; q_n^{\pi_n}) + \frac{3}{2} J(\pi_n, \mu_{n-1}; q_n^{\pi_n}) + \frac{3}{2} C_{\mathrm{B}}^{\mathrm{task}} J(\pi_n, \mu_{n-1}; v_n).$$

Substitute equation 98. □

D.2 Comparing task-only and optimal-value sensitivity

We use the optimal value defined in equation 72 and, for the remainder of this appendix, assume the optimal-return condition in equation 73.

Lemma D.2 (Task-only sensitivity versus optimal-value sensitivity)**.** *On the calibration event, for every n, t, j, and z,*

$$G_{n,t+1,j}^{\pi_n}(z) \le w_{t+1,j}^\star(z) + \frac{c_H H_{t+1}}{\sigma^2} \|\mu_{n-1}(z) - f^\star(z)\|_2 + \frac{1}{\sigma} \sqrt{H_{t+1} (P_{f^\star} |W_{n,t+1}^{\pi_n} - V_{t+1}^\star|)(z)}. \tag{100}$$

The continuation gaps along the true trajectory satisfy

$$\mathbb{E}_{\pi_n, f^\star} \left[\sum_{t=0}^{T-1} (P_{f^\star} |W_{n,t+1}^{\pi_n} - V_{t+1}^\star|)(z_{n,t}) \right] \le T \left[5 J(\pi_n, f^\star; q_n^{\pi_n}) + 6 C_{\mathrm{B}}^{\mathrm{task}} J(\pi_n, \mu_{n-1}; v_n) \right]. \tag{101}$$

Proof. The comparison equation 100 follows from Lemma A.2 with $U = W_{n,t+1}^{\pi_n}$, $V = V_{t+1}^\star$, $x = \mu_{n-1}(z)$, $y = f^\star(z)$, and $H = H_{t+1}$.

For the continuation bound, the triangle inequality and equation 95 give

$$|W_{n,t+1}^{\pi_n}(s) - V_{t+1}^\star(s)| \le \mathbb{E}_{\pi_n, f^\star} \left[\sum_{\ell=t+1}^{T-1} q_{n,\ell}^{\pi_n}(z_\ell) \,\middle|\, s_{t+1} = s \right] + V_{t+1}^\star(s) - V_{r,t+1}^{\pi_n, f^\star}(s).$$

As in the proof of Lemma C.2, the final difference is the expected future sum of nonnegative optimal Bellman gaps. Summing along the trajectory counts every one-step bound and gap at most T times, giving

$$\mathbb{E}_{\pi_n, f^\star} \left[\sum_{t} (P_{f^\star} |W_{n,t+1}^{\pi_n} - V_{t+1}^\star|)(z_{n,t}) \right] \le T \left[J(\pi_n, f^\star; q_n^{\pi_n}) + J(\pi^\star, f^\star; r) - J(\pi_n, f^\star; r) \right].$$

Substitute equation 99. □

D.3 Regret bound

Let $\mathcal{H}_N$ be the history collected by the task-only policies equation 94.

Theorem D.3 (Task-only regret from optimal-value relevance)**.** *Assume Assumptions 3.1, 3.2, and 3.4, the standing conditions of this section, and the GP conditions of Lemma A.13. Let $\beta_n \ge 1$ be the nondecreasing calibration coefficients from Assumption 3.4. With probability at least $1 - \delta$, the policies selected by equation 94 satisfy*

$$R_N = \widetilde{\mathcal{O}} \left(\beta_{N-1} \sqrt{N \, \widehat{\Gamma}_N^\star(k; \mathcal{H}_N)} + \beta_{N-1}^2 \Gamma_N(k) \right), \tag{102}$$

$$R_N = \widetilde{\mathcal{O}} \left(\beta_{N-1} \sqrt{N \, \Gamma_N^\star(k)} + \beta_{N-1}^2 \Gamma_N(k) \right). \tag{103}$$

Here $\widetilde{\mathcal{O}}$ suppresses logarithmic factors in N and $1/\delta$ and constants depending only on the fixed problem parameters, including the horizon T.

Proof. We drop the notation for conditioning episode-n expectations on $\mathcal{F}_{n-1}$ throughout the proof. Write $\beta = \beta_{N-1}$ and abbreviate

$$A^\star \triangleq \sum_{n=1}^{N} \mathbb{E}_{\pi_n, f^\star}\left[\sum_{t=0}^{T-1}\sum_{j=1}^{d_{\mathcal{S}}} w^\star_{t+1,j}(z_{n,t})\sigma_{n-1,j}(z_{n,t})\right],$$

$$B \triangleq \sum_{n=1}^{N} J(\pi_n, f^\star; q_n^{\pi_n}), \qquad V \triangleq \sum_{n=1}^{N} J(\pi_n, f^\star; u_{n-1}^2).$$

Let C denote a positive constant depending only on the fixed problem parameters; its value may change from line to line.

Transferring the correction. Apply Lemma A.9 with $(q, f, g, h) = (v_n, \mu_{n-1}, f^\star, f^\star)$. Since $v_n \le \sigma^2$, its constant is $C_v = 2T^2$. Calibration bounds the truncated squared discrepancy by v_n itself, so

$$J(\pi_n, \mu_{n-1}; v_n) \le (1 + \sqrt{C_v} + C_v) J(\pi_n, f^\star; v_n). \tag{104}$$

Because $v_n \le \beta_{n-1}^2 u_{n-1}^2$,

$$\sum_{n=1}^{N} J(\pi_n, \mu_{n-1}; v_n) \le (1 + \sqrt{C_v} + C_v)\beta^2 V. \tag{105}$$

Bounding regret. Summing equation 99 and using equation 105 gives

$$R_N \le 4B + C\beta^2 V. \tag{106}$$

Likewise, equation 101 gives

$$\sum_{n=1}^{N} \mathbb{E}_{\pi_n, f^\star}\left[\sum_t (P_{f^\star}|W^{\pi_n}_{n,t+1} - V^\star_{t+1}|)(z_{n,t})\right] \le T(5B + C\beta^2 V). \tag{107}$$

Multiply equation 100 by $\beta_{n-1}\sigma_{n-1,j}(z)$ and sum over coordinates, stages, and episodes. The definition of $q^\pi_{n,t}$, monotonicity of β_n, calibration, $\|\sigma_{n-1}(z)\|_1 \le \sqrt{d_{\mathcal{S}}} u_{n-1}(z)$, Cauchy–Schwarz, and equation 107 yield

$$B \le \beta A^\star + C\beta^2 V + C\beta\sqrt{V(B + \beta^2 V)}. \tag{108}$$

Splitting the square root and applying Young's inequality absorb $C\beta\sqrt{VB}$ into $B/2 + C\beta^2 V$, as in equation 88. Together with equation 106, this gives

$$R_N \le C(\beta A^\star + \beta^2 V). \tag{109}$$

Apply Lemma A.13 with $w = w^\star$, $\overline{w} = w_{\max}$, and $\eta = \delta/3$, and intersect with calibration. As in the proof of Theorem C.3, with probability at least $1 - \delta$ this yields

$$R_N \le C\left[\beta\sqrt{N\,\widehat{\Gamma}^\star_N(k; \mathcal{H}_N)} + \beta\log\frac{3}{\delta} + \beta^2\left(\Gamma_N(k) + \log\frac{3}{\delta}\right)\right].$$

Since $\beta \ge 1$, the fixed-kernel observation in Appendix A.6 absorbs the logarithmic terms and proves equation 102. Applying equation 46 directly to equation 109 proves equation 103, with constants depending only on the upper kernel and weight bounds and the fixed problem parameters. □

For the standard calibration of Chowdhury & Gopalan (2017), $\beta_{N-1}^2 = \widetilde{\mathcal{O}}(\Gamma_N(k))$, so the worst-case bound becomes $\widetilde{\mathcal{O}}(\sqrt{N\,\Gamma_N(k)\Gamma^\star_N(k)} + \Gamma_N(k)^2)$. A smaller valid calibration coefficient improves both terms in the theorem directly. Appendix H discusses settings where this is possible.

E Undirected Confidence Bonuses: A Tightened SOMBRL Analysis

The undirected confidence width equation 31 leads to SOMBRL's uncertainty reward u_{n-1}. We sharpen its regret bound by using the accumulated GP prediction error to control the bonus-return transfer.

E.1 Planning criterion and return transfers

We use the common setting of Appendix A.6 and the clipped squared prediction error ℓ_n from equation 49. SOMBRL uses the coefficient

$$\lambda_n^{\mathrm{SOMBRL}} \triangleq \frac{R_{\max} T}{\sigma}\, \beta_{n-1} \tag{110}$$

and selects

$$\pi_n \in \arg\max_{\pi \in \Pi} \left\{ J(\pi, \mu_{n-1}; r) + \lambda_n^{\mathrm{SOMBRL}} J(\pi, \mu_{n-1}; u_{n-1}) \right\}. \tag{111}$$

Lemma E.1 (Task- and bonus-return transfers)**.** *On the calibration event, for every episode n and every fixed or $\mathcal{F}_{n-1}$-measurable policy $\pi \in \Pi$,*

$$|J(\pi, f^\star; r) - J(\pi, \mu_{n-1}; r)| \le \lambda_n^{\mathrm{SOMBRL}} J(\pi, \mu_{n-1}; u_{n-1}). \tag{112}$$

For the uncertainty return, let

$$C_U \triangleq \max\left\{ \frac{2\sqrt{d_{\mathcal{S}} \sigma_{\max}}\, T^2}{\sigma^2},\, 1 \right\}. \tag{113}$$

Then

$$J(\pi, \mu_{n-1}; u_{n-1}) \le 2J(\pi, f^\star; u_{n-1}) + 2C_U J(\pi, f^\star; \ell_n). \tag{114}$$

Proof. Corollary A.7, with $q = r$ and $h = \mu_{n-1}$, gives equation 112 because calibration implies $\|\mu_{n-1}(z) - f^\star(z)\|_2 \le \beta_{n-1} u_{n-1}(z)$.

For the bonus return, apply equation 36 in Lemma A.9 with $q = u_{n-1}$, $f = \mu_{n-1}$, $g = f^\star$, and $h = f^\star$. By equation 52, its constant equation 37 is C_U. Applying $\sqrt{xy} \le (x+y)/2$ to the square-root term gives equation 114. □

E.2 Tightened regret bound

Theorem E.2 (Tightened SOMBRL regret)**.** *Under the common calibration setting and the assumptions of Lemma A.15, let $\beta_n \ge 1$ be the nondecreasing confidence multipliers from Assumption 3.4. Then, with probability at least $1 - \delta$, the policies selected by equation 111 satisfy*

$$R_N = \mathcal{O}\left(\beta_{N-1} \left[\sqrt{N\, \Gamma_N(k)} + \left(B^2 + \Gamma_N(k) + \log \frac{9 d_{\mathcal{S}}}{\delta} \right) \log \frac{9 N d_{\mathcal{S}}}{\delta} \right] \right), \tag{115}$$

where the hidden constant depends only on $T, d_{\mathcal{S}}, \sigma, \sigma_{\max}$, and $R_{\max}$. For fixed problem parameters and confidence level, this simplifies to

$$R_N = \widetilde{\mathcal{O}}\left(\beta_{N-1} \sqrt{N\, \Gamma_N(k)} \right). \tag{116}$$

Proof. On calibration, equation 112 makes the planning criterion an upper bound on each candidate's true return. Optimality of π_n and the same two-sided confidence bound give

$$J(\pi^\star, f^\star; r) - J(\pi_n, f^\star; r) \le 2\lambda_n^{\mathrm{SOMBRL}} J(\pi_n, \mu_{n-1}; u_{n-1}). \tag{117}$$

Define

$$S_N \triangleq \sum_{n=1}^{N} J(\pi_n, f^\star; u_{n-1}), \qquad E_N \triangleq \sum_{n=1}^{N} J(\pi_n, f^\star; \ell_n).$$

Sum equation 117 over episodes and apply the bonus-return transfer equation 114. Since $\lambda_n^{\mathrm{SOMBRL}} \leq (R_{\max}T/\sigma)\beta_{N-1}$, this gives

$$R_N \leq \frac{4R_{\max}T}{\sigma}\beta_{N-1}\left[S_N + C_U E_N\right]. \tag{118}$$

The posterior-width bound equation 44, with $\eta = \delta/3$, gives

$$S_N = \mathcal{O}\left(\sqrt{N\Gamma_N(k)} + \log\frac{3}{\delta}\right)$$

with probability at least $1 - \delta/3$. Apply Lemma A.15 with $\eta = \delta/3$ to control E_N. Substitution into equation 118 and a union bound with the calibration event prove equation 115.

For equation 116, the bounded kernel gives $\Gamma_N(k) \leq \frac{NT}{2}\log(1 + \sigma_{\max}/\sigma^2) = \mathcal{O}(N)$, so $\Gamma_N(k) = \mathcal{O}(\sqrt{N\Gamma_N(k)})$. The fixed-kernel observation in Appendix A.6 absorbs the remaining fixed terms up to logarithmic factors and covers $k \equiv 0$. □

For the standard confidence sequence of Chowdhury & Gopalan (2017), $\beta_{N-1}^2 = \widetilde{\mathcal{O}}(\Gamma_N(k))$ for fixed problem parameters and a nonzero kernel. Substituting into equation 116 gives $R_N = \widetilde{\mathcal{O}}(\sqrt{N}\,\Gamma_N(k))$. This has the same N- and information-gain exponents as the information-theoretic rate of Kakade et al. (2020), and improves the $\widetilde{\mathcal{O}}(\sqrt{N}\,\Gamma_N(k)^{3/2})$ scaling quoted for the earlier SOMBRL analysis in equation 10 by a factor $\sqrt{\Gamma_N(k)}$. Any tighter calibrated confidence sequence improves equation 116 through the factor β_{N-1}.

F When Value-Weighted Information Gain Is Smaller: A Split Bound

To highlight the potential benefit of value weighting, we first consider a single scalar regression problem on an input space $\mathcal{Z}$. Observations have Gaussian noise variance $\sigma^2 > 0$, and the learner uses a GP with kernel k, $k(z,z) \leq \sigma_{\max}$, updated after every observation. Write $v_{\ell-1}(z)$ for its posterior variance after the first $\ell - 1$ observations. For a fixed weight function $w : \mathcal{Z} \to [0,1]$, define

$$\gamma_m^w(k) \triangleq \sup_{z_1,\dots,z_m \in \mathcal{Z}} \frac{1}{2}\sum_{\ell=1}^{m} \log\left(1 + \frac{w(z_\ell)^2 v_{\ell-1}(z_\ell)}{\sigma^2}\right). \tag{119}$$

The posterior is the ordinary GP posterior: the weights reweight each marginal information gain contribution according to the sampled input z_ℓ, without affecting the GP posterior itself. We use normalized weights here to keep the comparison simple, and relate them to the paper's value gradients below.

F.1 Regional information gain and the split bound

For a nonempty region $A \subseteq \mathcal{Z}$ and a noise variance $\lambda > 0$, let

$$\gamma_m(A;\lambda) \triangleq \sup_{z_1,\dots,z_m \in A} \frac{1}{2}\log\det\left(I_m + \lambda^{-1}K_m\right), \qquad [K_m]_{ab} = k(z_a, z_b), \tag{120}$$

with $\gamma_0(A;\lambda) = 0$. This is ordinary maximum information gain restricted to A (Srinivas et al., 2010); repeated inputs are allowed. The kernel is fixed throughout and suppressed in this notation.

The following result bounds weighted information gain by a sum of regional information gains. Regions with smaller weights contribute less, either through a larger effective noise variance or an explicit squared-weight factor.

Proposition F.1 (Scalar split bound)**.** *Let $A_1, \dots, A_M$ partition $\mathcal{Z}$ into nonempty regions, and suppose $w(z) \leq c_i \leq 1$ on A_i, with $c_i \geq 0$. Then*

$$\gamma_m^w(k) \leq \max_{\substack{m_1,\dots,m_M \in \mathbb{Z}_{\geq 0} \\ m_1+\cdots+m_M = m}} \sum_{i=1}^{M} \gamma_{m_i}(A_i; \sigma^2/c_i^2), \tag{121}$$

where a summand with $c_i = 0$ is defined to be zero. Also, with $C_\sigma = 1 + \sigma_{\max}/\sigma^2$,

$$\gamma_m^w(k) \le C_\sigma \max_{\substack{m_1,\dots,m_M \in \mathbb{Z}_{\ge 0} \\ m_1+\cdots+m_M=m}} \sum_{i=1}^{M} c_i^2 \gamma_{m_i}(A_i; \sigma^2). \tag{122}$$

Proof. Fix a sequence of m inputs, with m_i visits to A_i. At each visit to A_i, discard all earlier observations outside A_i, and denote the resulting posterior variance by $v_{i,\ell-1}^{(\sigma^2)}$, where ℓ counts visits within this region. Discarding observations increases variance. For $c_i > 0$, further increase the noise variance to $\lambda_i = \sigma^2/c_i^2 \ge \sigma^2$. Thus, at the ℓth visit z to A_i, the corresponding summand in equation 119 is at most

$$\frac{1}{2}\log\left(1 + \frac{c_i^2 v_{i,\ell-1}^{(\sigma^2)}(z)}{\sigma^2}\right) = \frac{1}{2}\log\left(1 + \frac{v_{i,\ell-1}^{(\sigma^2)}(z)}{\lambda_i}\right) \le \frac{1}{2}\log\left(1 + \frac{v_{i,\ell-1}^{(\lambda_i)}(z)}{\lambda_i}\right).$$

Summing over the visits to A_i gives at most $\gamma_{m_i}(A_i; \lambda_i)$ by the sequential GP information identity. Regions with $c_i = 0$ contribute nothing. Sum over regions and maximize over the visit counts to obtain equation 121.

For the second bound, retain noise variance σ^2 and use

$$\log(1 + c_i^2 u) \le c_i^2 u \le C_\sigma c_i^2 \log(1+u), \qquad 0 \le u \le \sigma_{\max}/\sigma^2.$$

Apply this with $u = v_{i,\ell-1}^{(\sigma^2)}(z)/\sigma^2$ and use the same information identity within each region. □

The effective noise σ^2/c_i^2 is only a proof device. Since $c_i \le 1$, it is at least the learner's noise variance σ^2, so the corresponding regional posterior variance is no smaller than the posterior variance obtained with noise σ^2.

F.2 Intuition and comparison with unweighted information gain

High and low value sensitivity. Let A contain the value-sensitive inputs, with $w \le 1$ on A and $w \le \varepsilon$ on A^c, where $0 \le \varepsilon \le 1$ and both regions are nonempty. The proof above gives

$$\gamma_m^w(k) \le \max_{0 \le q \le m} \left\{\gamma_q(A; \sigma^2) + C_\sigma \varepsilon^2 \gamma_{m-q}(A^c; \sigma^2)\right\}, \tag{123}$$

where q is an integer. The relevant region may have a simple restricted kernel, while uncertainty elsewhere is discounted by ε^2. For $\varepsilon = 0$, only $\gamma_m(A; \sigma^2)$ remains, regardless of the kernel complexity outside A. The effect of a region's size depends on the kernel, as the linear example below shows.

An exact example of a reduction. For indicator weights $w = \mathbf{1}_A$, the comparison is exact:

$$\gamma_m^{\mathbf{1}_A}(k) = \gamma_m(A; \sigma^2). \tag{124}$$

Linear kernel and region size. Consider $k(z, z') = z^\top z'$ on the ball $\mathcal{Z} = \{z \in \mathbb{R}^d : \|z\|_2 \le R\}$. Let $A = \{z : \|z\|_2 \le r\}$, with $0 < r < R$, have weight one, and let the outer region A^c have weight $\varepsilon \in [0, 1]$. The finite-rank bound gives

$$\gamma_q(A; \lambda) \le \frac{d}{2}\log\left(1 + \frac{qr^2}{d\lambda}\right),$$

since $\sum_{\ell=1}^{q} z_\ell z_\ell^\top$ has rank at most d and trace at most qr^2. Applying equation 121 to the two regions therefore yields

$$\gamma_m^w(k) \le \max_{0 \le q \le m} \left\{\frac{d}{2}\log\left(1 + \frac{qr^2}{d\sigma^2}\right) + \frac{d}{2}\log\left(1 + \frac{(m-q)\varepsilon^2 R^2}{d\sigma^2}\right)\right\}. \tag{125}$$

The global upper bound, using weight one and radius R everywhere, is $\frac{d}{2}\log(1 + mR^2/(d\sigma^2))$. The split bound pairs the high weight with the smaller radius r, and the larger radius R with the low weight ε. For these concentric balls, $r^2/R^2 = (\mathrm{vol}(A)/\mathrm{vol}(\mathcal{Z}))^{2/d}$, so the role of regional volume is explicit. At a fixed sample budget, small r/R and ε can make the split bound much smaller. The split bound need not dominate the global bound for all parameter choices, so the tighter of the two may be used.

Squared-exponential and Matérn kernels. For isotropic squared-exponential and Matérn kernels, suppose the high-weight region A lies in a p-dimensional affine subspace of $\mathbb{R}^d$, with $p < d$. Distances are preserved in orthonormal coordinates on that subspace, so the restricted kernel is the same kernel on a p-dimensional domain. For compact A, fixed dimension, kernel parameters and noise variance $\lambda > 0$, applying Srinivas et al. (2012, Theorem 5) to an enclosing p-dimensional ball gives

$$\gamma_m(A;\lambda) = \begin{cases} \mathcal{O}\big((\log m)^{p+1}\big), & \text{squared-exponential,} \\ \mathcal{O}\big(m^{p(p+1)/(2\nu+p(p+1))}\log m\big), & \text{Matérn, } \nu > 1. \end{cases}$$

Thus, in equation 123, the high-weight region contributes according to its intrinsic dimension p, while the contribution from the low-weight region is multiplied by ε^2. When $\varepsilon = 0$, the weighted gain inherits the p-dimensional rate. For kernels applied to bounded features as in equation 38, the same argument applies when the feature image $\psi(A)$ lies in a p-dimensional affine subspace. The dimension condition is on $\psi(A)$, and compactness of the original state space is unnecessary.

Connection to the episodic regret bounds. The value-gradient weights $w^\star_{t+1,j}$ need not be at most one. Choose a common bound $W \geq 1$ on them and apply the scalar result to $w^\star_{t+1,j}/W$. Since $\log(1+W^2u) \leq W^2\log(1+u)$ for $u \geq 0$, this costs at most a factor W^2. For each stage and output coordinate, retain only observations from that stage in earlier episodes. The resulting sequential scalar posterior has at least as much variance as the learner's episode-frozen posterior, so

$$\Gamma^\star_N(k) \leq W^2 \sum_{t=0}^{T-1}\sum_{j=1}^{d_{\mathcal{S}}} \gamma_N^{w^\star_{t+1,j}/W}(k). \tag{126}$$

This normalization is only for the comparison; the bonuses and $\Gamma^\star_N(k)$ retain their original definitions. The ordinary gain is $\Gamma_N(k) = \gamma_{NT}(\mathcal{Z};\sigma^2)$. The next appendix gives a control problem where only one stage and output coordinate contribute to the weighted gain.

G A Noisy TV with Many Channels

We construct a two-step control problem with a TV that has D channels, derive its value sensitivities, and compare the resulting information-gain and regret bounds. The example separates the maximum gains $\Gamma^\star_N(k)$ and $\Gamma_N(k)$ while holding the state dimension and RKHS norm bound fixed.

Act on the task or watch the TV. Let $s = (x, y)$, where x describes task progress and y is the observed TV display. There are two task actions, $u \in \{-1, +1\}$, and D actions $\mathrm{TV}_1, \ldots, \mathrm{TV}_D$ that select a channel. With $T = 2$, independent noise $\epsilon_t \sim \mathcal{N}(0, \sigma^2 I_2)$, and $r(s,a) = (1+\sin x)/2$, set

$$f^\star(s,a) = \begin{cases} (\theta u, 0), & a = u \in \{-1,+1\}, \\ (-\pi/2, b_i \sin y), & a = \mathrm{TV}_i, \end{cases} \qquad 0 < |\theta| \leq \pi/4. \tag{127}$$

The coefficients $\theta, b_1, \ldots, b_D$ are unknown, with $\sum_i b_i^2 \leq B_0^2$ for a fixed B_0. Each episode starts at $x_0 = -\pi/2$ and $y_0 \sim \mathcal{N}(0, \sigma^2)$. A TV action sets the next task mean to $-\pi/2$ and the next display mean to $b_i \sin y$.

Optimal values. Since $T = 2$, the terminal value is $V_2^\star \equiv 0$. At the last decision, the Bellman recursion equation 72 therefore gives

$$V_1^\star(s) = \max_a \{r(s,a) + \mathbb{E}_\epsilon[V_2^\star(f^\star(s,a)+\epsilon)]\} = \frac{1+\sin x}{2},$$

because the reward does not depend on the action. For Gaussian noise, $\mathbb{E}\sin\epsilon_x = 0$ and $\mathbb{E}\cos\epsilon_x = e^{-\sigma^2/2}$, so

$$\overline{V}_1^\star(\xi) = \frac{1}{2} + \frac{1}{2}\mathbb{E}\sin(\xi_x + \epsilon_x) = \frac{1}{2} + c\sin\xi_x, \qquad c = \frac{1}{2}e^{-\sigma^2/2}.$$

At the first decision, the next task mean is θ, $-\theta$, or $-\pi/2$. Substituting these into the Bellman recursion yields

$$\begin{aligned} V_0^\star(s) &= \frac{1+\sin x}{2} + \frac{1}{2} + c\max\{\sin\theta, -\sin\theta, -1\} \\ &= \frac{1+\sin x}{2} + \frac{1}{2} + c\sin|\theta|. \end{aligned}$$

Thus the best first action is $u = \operatorname{sign}(\theta)$, while the last action is arbitrary. At $x_0 = -\pi/2$, the optimal return is $1/2 + c\sin|\theta|$; watching any channel first gives $1/2 - c$, a loss of $c(1 + \sin|\theta|)$. For the regret comparison, take Π to be the finite class of policies that choose a fixed action at each stage; it contains an optimal policy and the planning maxima are attained.

Where value is sensitive. Let $A = \mathcal{S} \times \{-1, +1\}$ be the task region in the notation of Appendix F. Differentiating the smoothed values gives

$$\nabla \overline{V}_1^\star(\xi) = (c\cos\xi_x, 0), \qquad \nabla \overline{V}_2^\star \equiv 0.$$

The weights in equation 77 evaluate the absolute derivatives at $\xi = f^\star(z)$. For task actions, $\cos(\theta u) = \cos\theta > 0$; for TV actions, $\cos(-\pi/2) = 0$. Hence

$$w_{1,x}^\star(z) = c\cos\theta\, \mathbf{1}_A(z), \qquad w_{1,y}^\star(z) = 0, \qquad w_{2,x}^\star(z) = w_{2,y}^\star(z) = 0. \tag{128}$$

A model that includes the TV inputs. The learner knows the kernel and an RKHS norm bound, and models both outputs without being given their structural zeros. Both scalar GP models use the linear kernel $k(z, z') = \phi(z)^\top \phi(z')$, with feature map

$$\phi(s, a) = \begin{cases} (u, 0, 0_D), & a = u \in \{-1, +1\}, \\ (0, 1, e_i \sin y)/\sqrt{2}, & a = \mathrm{TV}_i, \end{cases} \tag{129}$$

where e_i is the ith coordinate vector in $\mathbb{R}^D$. Writing $b = (b_1, \ldots, b_D)$, the true dynamics are linear combinations of these features:

$$f_x^\star(s, a) = \left\langle (\theta, -\pi/\sqrt{2}, 0_D), \phi(s, a) \right\rangle,$$
$$f_y^\star(s, a) = \left\langle (0, 0, \sqrt{2}b), \phi(s, a) \right\rangle.$$

For task actions these expressions give $(\theta u, 0)$. For a TV action, the factors of $\sqrt{2}$ compensate for the feature normalization, giving $(-\pi/2, b_i \sin y)$. The kernel has $k(z, z) \le 1$. Since the features span $\mathbb{R}^{D+2}$, the squared RKHS norms equal the squared Euclidean norms of the coefficient vectors:

$$\|f_x^\star\|_k^2 = \theta^2 + \pi^2/2, \qquad \|f_y^\star\|_k^2 = 2\sum_i b_i^2.$$

Hence the RKHS bound does not grow with D.

Proposition G.1 (Maximum information gains for the noisy TV)**.** *For every $N \ge D$,*

$$\Gamma_N^\star(k) \le \gamma_N(A; \sigma^2/c^2) = \frac{1}{2}\log\left(1 + \frac{Nc^2}{\sigma^2}\right),$$
$$\Gamma_N(k) \ge \frac{D}{2}\log\left(1 + \frac{N}{2D\sigma^2}\right). \tag{130}$$

In particular, for fixed $\sigma > 0$, $\Gamma_N^\star(k) = \mathcal{O}(\log(2 + N))$ whereas $\Gamma_N(k) = \Theta(D\log(1 + N/D))$.

Proof. Only the first stage and x coordinate have nonzero weights. Retain only earlier first-stage observations of that coordinate; discarding the other observations increases posterior variance. We can therefore apply the scalar split bound, Proposition F.1, to A, A^c, with weights at most $c \le 1$ on A and zero on A^c. On A, the kernel is $k(z, z') = uu'$, so the Gram matrix for N queries has one nonzero eigenvalue, equal to N. This gives the first bound.

For the ordinary gain, $T = 2$, so $\Gamma_N(k)$ maximizes information gain over at most $2N$ inputs. Set $m = \lfloor 2N/D \rfloor$ and choose m inputs for each TV channel, all with $y = \pi/2$ and distinct x-coordinates. This uses $mD \le 2N$ inputs. Since $\sin(\pi/2) = 1$ and the features do not depend on x, every selected input on channel i has feature vector $(0, 1, e_i)/\sqrt{2}$. For two selected inputs z and z' from channels i and j, respectively, their kernel entry is

$$k(z, z') = \frac{1 + e_i^\top e_j}{2} = \frac{1 + \mathbf{1}_{\{i=j\}}}{2}.$$

Thus the shared TV feature contributes $1/2$ to every entry of the Gram matrix. Each channel's own feature contributes an additional $1/2$ when both inputs belong to that channel. Grouping the m inputs from each channel together places these additional contributions in D diagonal $m \times m$ blocks, giving

$$K = \frac{1}{2}\mathbf{1}_{mD}\mathbf{1}_{mD}^\top + K_0, \qquad K_0 = \frac{1}{2}\operatorname{diag}(J_m, \ldots, J_m), \qquad J_m = \mathbf{1}_m\mathbf{1}_m^\top.$$

The first term is positive semidefinite, so $K \succeq K_0$. Each all-ones block J_m has one eigenvalue m and all others zero, so K_0 has D nonzero eigenvalues, each equal to $m/2$. The log-determinant expression for information gain now gives

$$\begin{aligned}\Gamma_N(k) &\geq \frac{1}{2}\log\det(I + \sigma^{-2}K) \geq \frac{1}{2}\log\det(I + \sigma^{-2}K_0) \\ &= \frac{D}{2}\log\Big(1 + \frac{m}{2\sigma^2}\Big) \geq \frac{D}{2}\log\left(1 + \frac{N}{2D\sigma^2}\right),\end{aligned}$$

where the last inequality uses $m = \lfloor 2N/D \rfloor \geq N/D$ for $N \geq D$.

Finally, the full kernel has rank at most $D + 2$ and diagonal at most one, so the rank–trace bound gives $\Gamma_N(k) \leq \frac{D+2}{2}\log(1 + \frac{2N}{(D+2)\sigma^2})$. □

The improvement in the regret bound. For the same calibration coefficient $\beta = \beta_{N-1}$, the C/D and SOMBRL bounds become, respectively,

$$R_N = \widetilde{\mathcal{O}}(\beta\sqrt{N} + \beta^2 D), \qquad \text{versus} \qquad \widetilde{\mathcal{O}}(\beta\sqrt{ND}). \tag{131}$$

With the standard choice $\beta = \widetilde{\mathcal{O}}(\sqrt{D})$, these give $\widetilde{\mathcal{O}}(\sqrt{ND} + D^2)$ and $\widetilde{\mathcal{O}}(D\sqrt{N})$. Here logarithms are suppressed, but powers of D are not. Relative to the unweighted scale $D\sqrt{N}$, the two terms in the value-directed bound have sizes $D^{-1/2}$ and $D/\sqrt{N}$, respectively. Thus, for many channels, the full bound improves when N is sufficiently larger than D^2, up to logarithmic factors. The stronger condition that N is sufficiently larger than D^3 makes the quadratic correction smaller than the value-directed leading term, exposing the full factor of order $\sqrt{D}$. These comparisons concern only regret upper bounds and do not by themselves imply that value-directed algorithms achieve the stated improvement over undirected algorithms.

H Can Calibration Depend on the Value-Weighted Information Gain?

The confidence multiplier and the accumulated posterior uncertainty play different roles in the regret bound. This appendix asks whether the confidence multiplier can depend on the value-weighted gain $\Gamma_N^\star(k)$ in place of the ordinary gain $\Gamma_N(k)$. We give a two-step example where these gains are separated, but even applying this replacement only to the leading term would imply a regret rate that no algorithm can attain. We then discuss improvements under additional structure.

H.1 The calibration question

The regret bounds in Appendices C and D give

$$R_N = \widetilde{\mathcal{O}}\left(\beta_{N-1}\sqrt{N\,\Gamma_N^\star(k)} + \beta_{N-1}^2\Gamma_N(k)\right). \tag{132}$$

The standard confidence choice has $\beta_{N-1}^2 = \widetilde{\mathcal{O}}(B^2 + \Gamma_N(k))$. Since prediction errors matter through their effect on policy returns, one may ask whether calibration can instead use $\beta_{N-1}^2 = \widetilde{\mathcal{O}}(B^2 + \Gamma_N^\star(k))$, uniformly over the RKHS ball. For fixed $B > 0$, applying this improvement only to the leading term of equation 132, while retaining the standard quadratic correction, would give

$$R_N = \widetilde{\mathcal{O}}\left(\sqrt{N\,\Gamma_N^\star(k)\big(B^2 + \Gamma_N^\star(k)\big)} + \Gamma_N(k)^2\right). \tag{133}$$

The following example rules out even this rate. It therefore also rules out replacing the confidence multiplier in both terms.

H.2 A LOWER BOUND WITH SEPARATED INFORMATION GAINS

The example is a two-step problem whose optimal first action matches an unknown vector of signs. Small value sensitivity suppresses the weighted gain without making these signs easy to learn.

Theorem H.1 (Value-weighted gain does not suffice for calibration)**.** *There exist a fixed bounded kernel k, a fixed reward, and a fixed RKHS norm bound, with $d_{\mathcal{S}} = d_{\mathcal{A}} = 1$, $T = 2$, $\sigma = 1$, and $R_{\max} = 1$, such that the following holds. For every learning algorithm and every integer $J \geq 1$, at $N_J = 2^{5J}$ some truth in the same RKHS ball satisfies*

$$\Gamma_N^\star(k) \leq C \log(2+N), \qquad c_0 N^{1/4} \leq \Gamma_N(k) \leq C N^{1/4} \quad \textit{for all } N \geq 16,$$
$$\mathbb{P}\Big(R_{N_J} \geq c N_J^{3/5}\Big) > \frac{1}{5}. \tag{134}$$

The constants $c, c_0, C > 0$ do not depend on J, the algorithm, or the truth. The learner may randomize and use both transition observations from each episode.

At any fixed failure probability below $1/5$, both terms in equation 133 are $\widetilde{\mathcal{O}}(\sqrt{N})$, contradicting the lower bound $\Omega(N^{3/5})$ along $N = N_J$. Even a kernel-dependent hidden constant cannot absorb the factor $N^{3/5}/\sqrt{N} = N^{1/10}$. Thus calibration based only on the value-weighted gain cannot give the proposed leading term for general kernels, even with the original quadratic correction. Note that this does not establish that the standard calibration factor is optimal.

Proof idea. We construct one fixed kernel from independent feature blocks and place the unknown dynamics in a block containing d hidden signs. The feature scales $\eta = d^{-1/2}$ and $\lambda = d^{-3/2}$ make value sensitivity small while each sign changes an observation mean by only $\mathcal{O}(\lambda/d)$. At $N = d^2/\lambda^2 = d^5$, a two-point testing argument yields regret $\Omega(N\eta\lambda) = \Omega(N^{3/5})$. We then bound the weighted and ordinary information gains separately.

Proof. **The kernel and the unknown dynamics.** For $j \geq 1$, let $d_j = 2^j$, $\eta_j = d_j^{-1/2}$, and $\lambda_j = d_j^{-3/2}$. Block j contains actions $u \in \{-1, 1\}^{d_j}$ and has its own $(d_j + 1)$-dimensional feature space. Label each pair (j, u) by a distinct positive integer; throughout, (j, u) denotes this scalar action. Include a dummy action 0, so $\mathcal{A} \subset \mathbb{R}$. Define the feature map $\phi(z) = (\phi_1(z), \phi_2(z), \ldots)$ by

$$\phi_j(s, (j, u)) = (\eta_j, \lambda_j u/\sqrt{d_j}), \qquad \phi_b(s, (j, u)) = 0 \quad (b \neq j), \qquad \phi(s, 0) = 0.$$

Thus an action activates only its own block. The kernel $k(z, z') = \langle \phi(z), \phi(z') \rangle$ is

$$k((s, (j, u)), (s', (j', v))) = \mathbf{1}\{j = j'\} \left(\eta_j^2 + \frac{\lambda_j^2}{d_j} u^\top v\right), \qquad k((s, 0), z) = 0. \tag{135}$$

It ignores the state and satisfies $k(z, z) = d_j^{-1} + d_j^{-3} \leq 1$. Because this kernel is linear in ϕ, every square-summable coefficient vector θ defines an RKHS function $f_\theta(z) = \langle \theta, \phi(z) \rangle$ with $\|f_\theta\|_{\mathcal{H}_k} \leq \|\theta\|_2$.

Fix an active block J, write $d = d_J$, $\eta = \eta_J$, $\lambda = \lambda_J$, and set $\kappa = 1/16$. For an unknown $\tau \in \{-1, 1\}^d$, choose θ_τ with block J equal to $\kappa(3, \tau/\sqrt{d})$ and all other blocks zero. The dynamics are

$$f_\tau(z) = \langle \theta_\tau, \phi(z) \rangle = \begin{cases} \kappa\left(3\eta + \dfrac{\lambda}{d} \tau^\top u\right), & z = (s, (J, u)), \\ 0, & \text{otherwise.} \end{cases} \tag{136}$$

Their RKHS norms satisfy

$$\|f_\tau\|_{\mathcal{H}_k}^2 \leq \|\theta_\tau\|_2^2 = \kappa^2 \left(9 + \frac{\|\tau\|_2^2}{d}\right) = 10\kappa^2.$$

Thus every dynamics function f_τ lies in the same RKHS ball with $B = \kappa\sqrt{10}$. Since $\lambda \leq \eta$, the active means lie in $[2\kappa\eta, 4\kappa\eta] \subset (0, \pi/2)$, using $4\kappa\eta \leq 1/(4\sqrt{2}) < \pi/2$.

Values, sensitivities, and action gaps. Each episode starts at $s_0 = 0$, the noise is standard Gaussian, and $r(s, a) = (1 - \cos s)/2$. Write $q = e^{-1/2}$. Since the last action does not affect its immediate reward and $V_2^\star = 0$,

$$V_1^\star(s) = \frac{1 - \cos s}{2}, \qquad \overline{V}_1^\star(y) = \frac{1 - q\cos y}{2}. \tag{137}$$

Starting from $s_0 = 0$, the first reward is zero, so the expected return of first action a is $\overline{V}_1^\star(f_\tau(0, a))$, regardless of the last action. For $a = (J, u)$, we have $\tau^\top u \le d$, with equality at $u = \tau$. Thus (J, τ) has the largest transition mean $\kappa(3\eta + \lambda)$; actions outside block J have mean zero. The smoothed value is increasing on $[0, 4\kappa\eta]$, so this action also maximizes expected return. Let $\ell_\tau(a)$ denote the difference between this optimal return and the return of action a. This is the expected-return regret of an episode whose first action is a. The gap bound below converts incorrect signs into regret.

For a full input $z = (s, a)$, the weight in equation 16 is the absolute derivative of the smoothed continuation evaluated at the transition mean. Differentiating gives

$$\begin{aligned} w_1^\star(z) &= \left|(\overline{V}_1^\star)'(f_\tau(z))\right| = \frac{q}{2}|\sin f_\tau(z)| \\ &\le \frac{q}{2}|f_\tau(z)| \le 2q\kappa\eta\,\mathbf{1}\{a \text{ in block } J\}. \end{aligned} \tag{138}$$

The inequalities use $|\sin x| \le |x|$ and the bounds on f_τ, which is zero outside the active block. Since $\overline{V}_2^\star = 0$, the final-transition weight is $w_2^\star(z) = 0$.

On $[2\kappa\eta, 4\kappa\eta]$, the derivative $q \sin x/2$ is at least $2q\kappa\eta/\pi$, using $\sin x \ge 2x/\pi$ for $x \in [0, \pi/2]$. For $a = (J, u)$, the gap is the difference in smoothed values at $f_\tau(0, a)$ and $\kappa(3\eta + \lambda)$. The fundamental theorem of calculus gives the first equality below; replacing the derivative by its lower bound gives the inequality:

$$\begin{aligned} \ell_\tau(a) &= \int_{f_\tau(0,a)}^{\kappa(3\eta+\lambda)} \frac{q}{2}\sin x\, dx \\ &\ge \frac{2q\kappa\eta}{\pi}\left[\kappa(3\eta + \lambda) - f_\tau(0, a)\right] \\ &= \frac{2q\kappa^2\eta\lambda}{\pi d}\sum_{i=1}^{d}(1 - \tau_i u_i). \end{aligned} \tag{139}$$

The last equality substitutes the dynamics from equation 136. For later use, define $u(a) = u$ if $a = (J, u)$ and $u(a) = 0$ otherwise. An action outside block J has mean zero, so its gap is at least the increase in smoothed value from $3\kappa\eta$ to $\kappa(3\eta + \lambda)$. The same derivative bound gives $\ell_\tau(a) \ge (2q\kappa\eta/\pi)\kappa\lambda = 2q\kappa^2\eta\lambda/\pi$, exactly the right-hand side of equation 139 at $u = 0$.

Why the signs cannot be learned quickly. Set $N = d^2/\lambda^2 = d^5$. Flipping one sign of τ changes the observation mean at any action by at most $2\kappa\lambda/d = 2\kappa/\sqrt{N}$. Fix a coordinate i and the other signs τ_{-i}. Let P_+ and P_- be the laws of the learner's full history when $\tau_i = +1$ and $\tau_i = -1$, respectively. There are at most $2N$ observations, including the transition after the final reward. Conditional on the same past history, the learner uses the same action-selection distribution under both truths, so only the transition laws contribute to the KL chain rule. At input (s, a), these laws are $\mathcal{N}(m_+, 1)$ and $\mathcal{N}(m_-, 1)$. Their unit noise variance gives

$$\mathrm{KL}(\mathcal{N}(m_+, 1)\,\|\,\mathcal{N}(m_-, 1)) = \frac{(m_+ - m_-)^2}{2} \le \frac{(2\kappa/\sqrt{N})^2}{2}.$$

Summing these conditional contributions and applying Pinsker's inequality gives

$$\mathrm{KL}(P_+\|P_-) \le 2N\frac{(2\kappa/\sqrt{N})^2}{2} = 4\kappa^2, \qquad \mathrm{TV}(P_+, P_-) \le \sqrt{2}\kappa. \tag{140}$$

The same bound holds for each history prefix, including the randomization used to select the next action. Let $a_{n,0}$ be the learner's first action in episode n, and write $U_{n,i} = u_i(a_{n,0})$. This is the ith sign encoded by the chosen action if it is in block J, and zero otherwise. Since $U_{n,i} \in [-1, 1]$, the total-variation bound in equation 140 therefore gives

$$|\mathbb{E}_+ U_{n,i} - \mathbb{E}_- U_{n,i}| \le 2\,\mathrm{TV}(P_+, P_-) \le 2\sqrt{2}\kappa.$$

Substituting this into the average sign mismatch yields

$$\frac{1}{2}\left(\mathbb{E}_+[1-U_{n,i}]+\mathbb{E}_-[1+U_{n,i}]\right)=1-\frac{\mathbb{E}_+U_{n,i}-\mathbb{E}_-U_{n,i}}{2}\geq 1-\sqrt{2}\kappa.$$

Averaging over the 2^{d-1} choices of τ_{-i} pairs all 2^d truths:

$$2^{-d}\sum_{\tau}\mathbb{E}_\tau[1-\tau_iU_{n,i}]=2^{-(d-1)}\sum_{\tau_{-i}}\frac{\mathbb{E}_+[1-U_{n,i}]+\mathbb{E}_-[1+U_{n,i}]}{2}\geq 1-\sqrt{2}\kappa.$$

For any learning history, average equation 139 over each selected policy's first action. The regret definition in equation 4 then gives

$$\begin{aligned}R_N&=\sum_{n=1}^{N}\mathbb{E}_{a\sim\pi_n(\cdot|0)}[\ell_\tau(a)]\\&\geq\frac{2q\kappa^2\eta\lambda}{\pi d}\sum_{n=1}^{N}\mathbb{E}_{a\sim\pi_n(\cdot|0)}\left[\sum_{i=1}^{d}(1-\tau_iu_i(a))\right]=:S_N.\end{aligned}$$

Conditional on the history before episode n and the selected policy, $a_{n,0}$ has distribution $\pi_n(\cdot\mid 0)$. Thus the tower rule gives

$$\mathbb{E}_\tau\left[\mathbb{E}_{a\sim\pi_n(\cdot|0)}[1-\tau_iu_i(a)]\right]=\mathbb{E}_\tau[1-\tau_iU_{n,i}].$$

Combining this identity with the average over truths, and summing over n,i,

$$\begin{aligned}2^{-d}\sum_{\tau}\mathbb{E}_\tau S_N&=\frac{2q\kappa^2\eta\lambda}{\pi d}\sum_{n=1}^{N}\sum_{i=1}^{d}2^{-d}\sum_{\tau}\mathbb{E}_\tau[1-\tau_iU_{n,i}]\\&\geq\frac{2q\kappa^2\eta\lambda}{\pi d}Nd(1-\sqrt{2}\kappa).\end{aligned}$$

At least one truth τ has expectation at least this average. For that truth,

$$\mathbb{E}_\tau S_N\geq c_LN\eta\lambda,\qquad 0\leq S_N\leq\frac{4q\kappa^2}{\pi}N\eta\lambda,\qquad c_L=\frac{2q\kappa^2}{\pi}(1-\sqrt{2}\kappa).\tag{141}$$

The upper bound follows from $\sum_i(1-\tau_iu_i)\leq 2d$. For a nonnegative variable $X\leq M$, $\mathbb{E}X\geq L$ implies $\mathbb{P}(X\geq L/2)\geq L/(2M)$. Applying this to S_N gives

$$\mathbb{P}_\tau\Big(R_N\geq\frac{c_L}{2}N\eta\lambda\Big)\geq\frac{1-\sqrt{2}\kappa}{4}>\frac{1}{5}.\tag{142}$$

At $N_J=d_J^5$, we have $N_J\eta_J\lambda_J=d_J^3=N_J^{3/5}$. The argument even allows the learner to know which block is active.

The value-weighted information gain. Only first transitions in block J have nonzero weights. Discard all other observations. This increases the posterior variance at each retained query, and leaves an ordinary sequential regression problem on that block with at most N inputs $z_i=(s_i,a_i)$. Denote these variances by p_i. They satisfy $0\leq p_i\leq\eta^2+\lambda^2\leq 2\eta^2\leq 1$. By equation 138 and $p\leq 2\log(1+p)$ on $[0,1]$,

$$\begin{aligned}\widehat{\Gamma}_N^\star(k;\mathcal{H}_N)&\leq\frac{1}{2}\sum_i(w_1^\star(z_i))^2p_i\leq 4q^2\kappa^2\eta^2\sum_i\log(1+p_i)\\&\leq 4q^2\kappa^2\eta^2(d+1)\log\left(1+\frac{2N\eta^2}{d+1}\right)\leq C\log(2+N).\end{aligned}\tag{143}$$

The second line uses the sequential log-determinant identity and the rank–trace bound on block J. The squared-sensitivity factor η^2 cancels the growing rank: $\eta^2(d+1)=(d+1)/d\leq 3/2$, uniformly in J. Taking the supremum over histories proves the weighted-gain bound.

The ordinary information gain. For any sequence of at most $2N$ inputs, the Gram matrix separates into blocks. Within block j, write it as the sum of the constant-feature and sign-feature Gram matrices.

The first has rank one and trace at most $2N\eta_j^2$. The second has rank at most d_j and trace at most $2N\lambda_j^2$. For positive semidefinite A, B,

$$\det(I + A + B) = \det(I + A)\det\Big(I + B^{1/2}(I + A)^{-1}B^{1/2}\Big) \le \det(I + A)\det(I + B),$$

where the inequality uses $(I + A)^{-1} \preceq I$. Combining this inequality with the rank–trace bound gives

$$\Gamma_N(k) \le \frac{1}{2}\sum_{j\ge1}\log(1 + 2N/d_j) + \frac{1}{2}\sum_{j\ge1} d_j \log(1 + 2N/d_j^4). \tag{144}$$

The first sum is $\mathcal{O}(\log^2(2 + N))$: there are $\mathcal{O}(\log N)$ blocks with $d_j \le 2N$, each contributing at most $\log(1+2N)$, and the remaining terms sum to at most $2N\sum_{d_j>2N} d_j^{-1} = \mathcal{O}(1)$. For the second sum, choose a dyadic D with $(2N)^{1/4}/2 < D \le (2N)^{1/4}$. For $d_j \le D$, write $d_j = 2^{-r}D$ and use $\log(1 + 2N/d_j^4) \le C(1 + r)$. For $d_j > D$, use $\log(1 + x) \le x$. Thus

$$\sum_{j\ge1} d_j \log(1 + 2N/d_j^4) \le CD\sum_{r\ge0} 2^{-r}(1 + r) + 2N\sum_{d_j>D} d_j^{-3} \le CN^{1/4}.$$

Since $\log^2(2 + N) = \mathcal{O}(N^{1/4})$, this proves the upper bound.

For the lower bound, choose a block with $N^{1/4}/2 < d_j \le N^{1/4}$, which need not be the active block: $\Gamma_N(k)$ maximizes over all inputs. Take d_j mutually orthogonal sign vectors, the rows of a Hadamard matrix. Such matrices exist for powers of two by the recursion $H_{2d} = \left(\begin{smallmatrix} H_d & H_d \\ H_d & -H_d \end{smallmatrix}\right)$ starting from $H_1 = (1)$. Repeat each selected action d_j^3 times at distinct states, which leaves its features unchanged, using $d_j^4 \le N$ queries. The sign features alone give d_j nonzero Gram eigenvalues $d_j^3\lambda_j^2 = 1$. Adding the constant feature can only increase the log determinant, so

$$\Gamma_N(k) \ge \frac{d_j}{2}\log 2 \ge \frac{\log 2}{4}N^{1/4}. \qquad \square$$

The example shows why small optimal-value derivatives do not by themselves justify a small calibration coefficient: they reduce $\Gamma_N^\star(k)$, but do not remove the need to identify the unknown signs. The lower bound applies to every algorithm, including batched algorithms. The finite-domain confidence bound of Li & Scarlett (2022, Corollary 1) contains $\sqrt{\log M}$ for M candidate inputs. On the active block, $M = 2^d$ gives a $\sqrt{d}$ factor. The blockwise improvement below is also consistent with the example: it requires small feature blocks, whereas these blocks have increasing dimension.

Remark (blockwise calibration). A smaller confidence multiplier follows from Abbasi-Yadkori et al. (2011, Theorem 2, Confidence Ellipsoid) when the kernel has independent feature blocks. Specifically, suppose $k(z, z') = \sum_{b=1}^K \langle\phi_b(z), \phi_b(z')\rangle$, with $\phi_b(z) \in \mathbb{R}^{r_b}$, $r_b \le r$, at most one nonzero block at each input, and $k(z, z) \le \sigma_{\max}$. Under the common coordinatewise RKHS norm bound B and Gaussian noise variance σ^2, apply their confidence ellipsoid separately to each block, normalizing features and responses by σ and using unit regularization. A union bound over blocks and output coordinates gives simultaneous coordinatewise GP confidence for adaptive inputs and all observation counts m, with failure probability at most $\delta/3$, using

$$\beta_m^{\mathrm{blk}} = \max\left\{1, B + \sqrt{r\log\Big(1 + \frac{m\sigma_{\max}}{r\sigma^2}\Big) + 2\log\frac{3Kd_{\mathcal{S}}}{\delta}}\right\}.$$

Observations update only the feature block they activate. Total information gain sums the contributions of all blocks, whereas each prediction involves only its active block. The confidence cost therefore depends on the largest block dimension and only logarithmically on K.

This can be substantially smaller than the generic $\widetilde{\mathcal{O}}(\sqrt{\Gamma_N(k)})$ multiplier when uncertainty is spread across many small blocks. For bounded r, B, K growing at most polynomially in N, and fixed remaining problem parameters, $(\beta_{(N-1)T}^{\mathrm{blk}})^2 = \widetilde{\mathcal{O}}(1)$. Using $\beta_{n-1} = \beta_{(n-1)T}^{\mathrm{blk}}$ in equation 84 or equation 103, under the corresponding planning assumptions, therefore yields $R_N = \widetilde{\mathcal{O}}(\sqrt{N\Gamma_N^\star(k)} + \Gamma_N(k))$.

Remark (batching and improved calibration). Batched bandit algorithms can use past observations to guide future decisions while retaining sharper confidence bounds for independent regression designs. In the batched pure-exploration algorithm of Li & Scarlett (2022, Algorithm 1, Corollary 1, and Theorem 1), earlier observations determine which actions remain plausible maximizers. Within a batch, maximum-variance sampling selects inputs without using that batch's labels, and each GP fit uses only that batch's observations. The surviving action set carries learning between batches. For a finite input domain of size M, at most E batches, RKHS norm bound B, and Gaussian noise with matching GP regularization, simultaneous confidence uses the multiplier $B+\sqrt{2\log(2ME/\delta)}$. Their exploration analysis yields cumulative bandit regret $\widetilde{\mathcal{O}}([B+\sqrt{\log(M/\delta)}]\sqrt{N\gamma_N(k)})$, where $\gamma_N(k)$ is the scalar information gain from N queries. Thus calibration need not introduce an additional $\sqrt{\gamma_N(k)}$ factor. Extending to the RL setting is left as future work.

I Experiment Details

Implementation. We implement each exploration strategy using MBPO (Janner et al., 2019) for state-based control and DreamerV3 (Hafner et al., 2025) for visual control. Within each setting, methods share the base architecture, exploration coefficient, dynamics model, and training schedule. For state-based control, an ensemble of five models predicts state differences and rewards. Each model uses two hidden layers of width 256 except the humanoid and quadruped experiments which use width 1024, following SOMBRL (Sukhija et al., 2025b), and a fixed randomized prior function (Osband et al., 2018). We optimize the actor, critic, and temperature parameters with Adam and the dynamics model with AdamW, all at learning rate 3×10^{-4}. The implementation builds on JAXRL (Kostrikov, 2021), with an intrinsic reward computation similar to MaxInfoRL (Sukhija et al., 2025a). For visual control, we use the model with 12 million parameters from the official DreamerV3 implementation and add sixteen two-layer disagreement heads of width 512, following SOMBRL (Sukhija et al., 2025b). Each head predicts next-step stochastic-latent categorical probabilities and symlog-transformed rewards from the current latent state and action. The heads are trained on posterior transitions using categorical cross-entropy and reward regression. Code for our approach and baselines is provided in the supplementary materials.

Value-directed bonus. Both implementations compute the value-directed bonus from a critic trained only on the task reward. The state-based implementation maintains five independent extrinsic critics and five augmented critics trained on the task reward plus exploration bonus. The actor uses the augmented value, intrinsic bonus, and SAC entropy term. DreamerV3 instead learns separate extrinsic and intrinsic value critics and combines their independently normalized advantages in the actor objective. For VDO-grad, the state-based implementation evaluates the extrinsic-value gradient at the average predicted next state and combines its coordinate sensitivities with the corresponding predictive uncertainties. DreamerV3 differentiates the extrinsic value with respect to the predicted next stochastic latent state and uses $\sqrt{g^\top \Sigma g}$, where g is this gradient and Σ is the ensemble covariance of the next latent predictions, including correlations across dimensions. In all cases, value gradients are computed by automatic differentiation through the learned value function.

Hyperparameters. For each task, we select the exploration coefficient $\lambda \in \{0.05, 0.1, 0.2\}$ that gives the highest SOMBRL performance and use it for all exploration methods. We hold λ fixed during training; SOMBRL's autotuning procedure (Sukhija et al., 2025b) reduced the bonus too quickly on challenging tasks. The reward-disagreement coefficient is fixed at 0.3 and controls the contribution of reward disagreement relative to next-state or latent disagreement. VDO-MBRL starts with the unweighted SOMBRL bonus during a task-specific warmup period, allowing the extrinsic critic to learn before its value estimates direct exploration. We stabilize the directed bonus with running-mean normalization and clipping, and retain uncertainty-based annealing as model disagreement decreases. Table 1 lists the exploration coefficients and warmup durations.

Computational Resources. In Table 2, we give the average wall-clock time required for training for each environment and method in addition to the slowdown relative to the corresponding backbone in parentheses. Note that the absolute runtime varies across environments because of differences in simulation, rendering, and number of total environment steps, so the relative slowdown provides the clearest comparison of computational overhead. The Dreamer-based version of VDO-grad induces

Table 1: Hyperparameters used across experiments.

Setting	Tasks	λ	VDO warmup
State noisy TV	PointMaze Large, Corridor, Lattice, Ring, Switchback (all distractor dimensions)	0.1	50k[†]
Vision noisy TV	FourRooms (all TV settings)	0.05	200k[‡]
State DMC	Cartpole Swingup Sparse, Cheetah Run, Finger Spin, Finger Turn Hard, Quadruped Run, Quadruped Walk, Reacher Hard, Walker Run	0.1	50k
	Humanoid Stand, Humanoid Walk, Humanoid Run	0.05	50k
	Hopper Hop	0.2	50k
Vision DMC	Hopper Hop	0.1	50k
	Finger Spin	0.2	50k
Atari	Enduro, Robotank	0.1	1M

[†]VDO-grad uses a 100k warmup in the state noisy TV experiments.
[‡]VDO-grad uses a 300k warmup in the vision noisy TV experiments.

similar computational overhead to SOMBRL since it only requires one extra critic gradient per imagined state. VDO-spread is more expensive because it evaluates the task critic many times per imagined state, specifically estimating each ensemble member's expected next-state value with $S = 64$ Monte Carlo samples in the vision implementation. We set S conservatively high, however small-scale experiments suggest that lower S would suffice. All experiments use a single RTX 2080 Ti GPU.

Table 2: Mean wall-clock training time in hours and runtime compared to the backbone (MBPO for state-based tasks, DreamerV3 for vision-based tasks).

Task	Backbone	SOMBRL	VDO-grad	VDO-spread
State-based noisy TV				
PointMaze Large	0.78	0.90 (1.15×)	1.21 (1.56×)	1.29 (1.66×)
PointMaze Lattice	1.46	1.60 (1.10×)	2.25 (1.54×)	2.44 (1.67×)
PointMaze Corridor	1.50	1.67 (1.11×)	2.31 (1.54×)	2.47 (1.64×)
PointMaze Ring	1.45	1.62 (1.12×)	2.27 (1.57×)	2.42 (1.67×)
PointMaze Switchback	1.47	1.66 (1.13×)	2.27 (1.54×)	2.46 (1.67×)
Vision-based noisy TV				
MiniWorld no TV	3.99	5.12 (1.28×)	5.22 (1.31×)	7.48 (1.87×)
MiniWorld deterministic TV	4.28	5.46 (1.27×)	5.58 (1.30×)	7.73 (1.81×)
MiniWorld stochastic TV	4.30	5.45 (1.27×)	5.60 (1.30×)	7.74 (1.80×)
State DMC				
Cartpole Swingup Sparse	0.34	0.39 (1.15×)	0.53 (1.56×)	0.57 (1.68×)
Cheetah Run	0.68	0.76 (1.12×)	1.00 (1.47×)	1.06 (1.56×)
Finger Spin	0.32	0.40 (1.25×)	0.52 (1.62×)	0.54 (1.69×)
Finger Turn Hard	0.72	0.78 (1.08×)	1.02 (1.42×)	1.15 (1.60×)
Hopper Hop	0.70	0.83 (1.19×)	1.06 (1.51×)	1.09 (1.56×)
Humanoid Run	6.82	7.85 (1.15×)	10.35 (1.52×)	12.83 (1.88×)
Humanoid Stand	6.90	7.99 (1.16×)	10.62 (1.54×)	12.97 (1.88×)
Humanoid Walk	7.21	7.93 (1.10×)	10.00 (1.39×)	13.35 (1.85×)
Quadruped Run	6.76	7.74 (1.14×)	10.41 (1.54×)	12.93 (1.91×)
Quadruped Walk	6.69	7.72 (1.15×)	10.10 (1.51×)	12.90 (1.93×)
Reacher Hard	0.33	0.38 (1.15×)	0.52 (1.58×)	0.55 (1.67×)
Walker Run	0.78	0.87 (1.12×)	1.08 (1.38×)	1.16 (1.49×)
Vision DMC				
Hopper Hop	58.3	78.4 (1.35×)	79.6 (1.37×)	119.7 (2.05×)
Finger Spin	14.3	19.7 (1.38×)	19.4 (1.36×)	28.5 (1.99×)
Atari				
Robotank	55.2	77.0 (1.40×)	77.1 (1.40×)	109.1 (1.98×)
Enduro	55.9	75.8 (1.36×)	77.4 (1.39×)	99.7 (1.78×)

J Additional Experiments

In this section, we provide additional experiments that extend the noisy TV analysis across harder maze layouts and TV placements, visualize representative behavior in the MiniWorld setting, and report results on standard state and vision-based control benchmarks.

State-based noisy TV. We extend the state-based noisy TV experiment from Section 6 to four more challenging PointMaze layouts using $d = 64$ distractor dimensions, with results shown in Figure 2 for layouts visualized in Figure 6. Each environment uses the same noisy TV construction as PointMaze Large with the TV cell placed along the route 5 cells from the start. Across the Lattice, Corridor, Ring, and Switchback layouts, VDO-grad and VDO-spread consistently achieve higher returns and lower TV occupancy than SOMBRL. Unlike PointMaze Large, these layouts are difficult for MBPO to solve reliably and the stronger performance of VDO-MBRL therefore indicates that its robustness does not come simply from suppressing uncertainty-driven exploration.

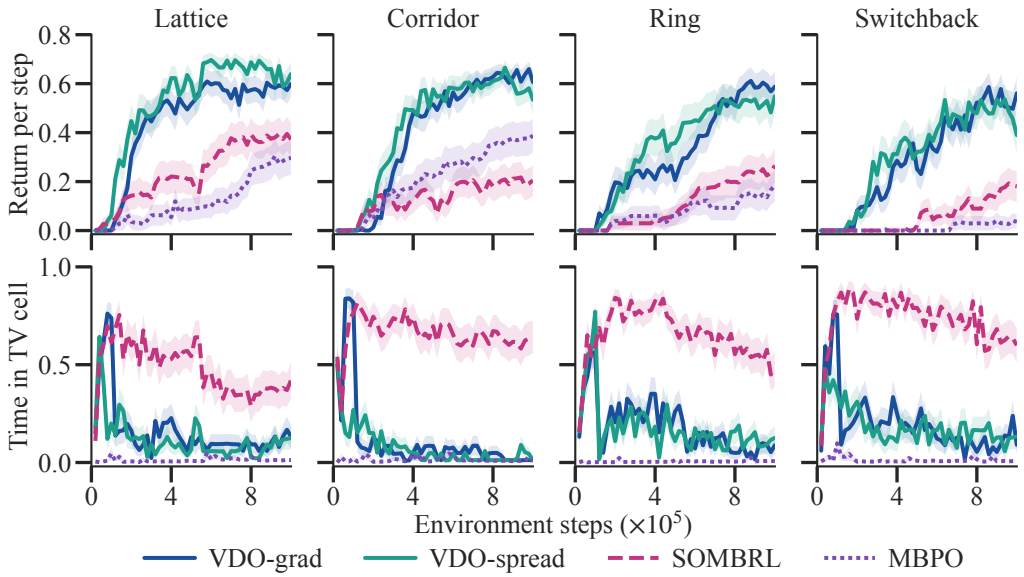


Figure 5: *Top:* Return per step over 25 seeds on four new PointMaze layouts with a noisy TV using $d = 64$ distractor dimensions. *Bottom:* Corresponding curves showing the TV proximity rates.

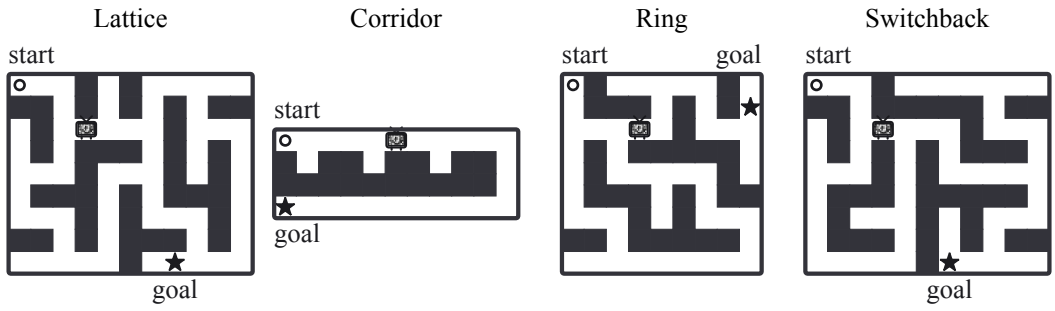


Figure 6: Maze layouts used in Figure 5. Circles mark the start, stars mark the goal, and TV icons mark the distractor cell.

TV placement. We also study the dependence of the state-based noisy TV results on the TV location by placing the TV at different four positions in PointMaze Large, ranging from close to the start to locations farther away or off the direct route to the goal. The results are given in Figure 7 and exact placements are shown in Figure 8. We find that VDO-MBRL maintains lower TV occupancy and generally higher returns than SOMBRL across placements, indicating that the observed failure mode is not specific to a single TV position. The placement does affect the strength of the distraction, with SOMBRL visiting the TV much less frequently when it is 11 cells from the start and more often when the TV is closer to the start. Note that the 5 cells from start configuration shown is the one used for the PointMaze Large setup in Figure 2.

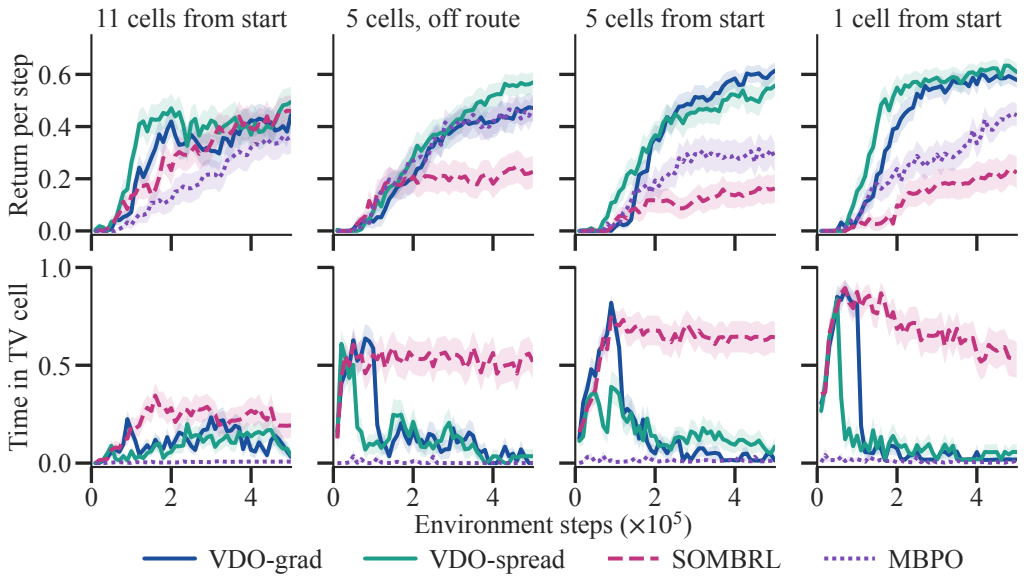


Figure 7: *Top:* Return per step over 25 seeds on PointMaze large with a noisy TV using $d = 64$ distractor dimensions placed at various locations. *Bottom:* Corresponding curves showing the TV proximity rates.

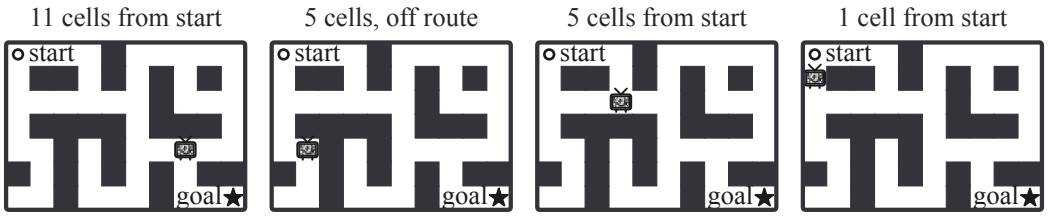


Figure 8: Visualization of the TV placements for Figure 7 in PointMaze Large. Circles mark the start, stars mark the goal, and TV icons mark the distractor cell.

Vision-based noisy TV rollouts. In Figure 9, we give top-down visualizations of rollouts from agents traine in the MiniWorld FourRooms environment with a stochastic TV. This illustrates how SOMBRL often remains near the TV and views it during the episode while VDO-spread spends less time watching the screen and instead progresses toward the goal. These trajectories qualitatively show the difference in screen visibility that is measured quantitatively in Figure 3.

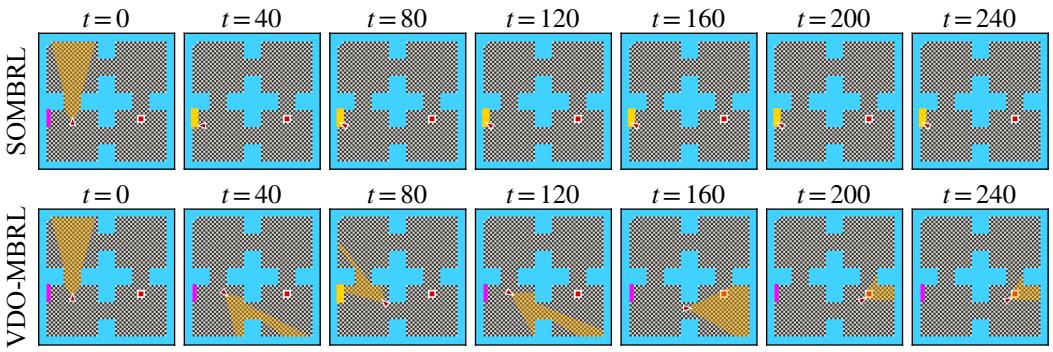


Figure 9: *Top:* Top-down view of a training episode for SOMBRL with the stochastic TV in MiniWorld FourRooms. *Bottom:* Corresponding view for VDO-spread. The arrow marks the agent and its heading, the shaded wedge show the agent's field of view, the magenta bar is the TV screen, and the red square is the goal.

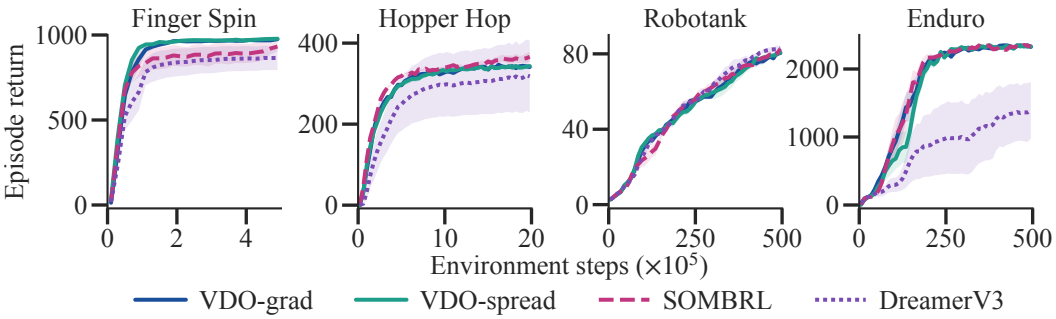


Figure 10: Learning curves over 5 seeds for vision-based tasks from DMC and Atari using DreamerV3 as the base algorithm.

Standard control benchmarks. To test value-directed exploration on unmodified control benchmarks, we evaluate on two vision-based DeepMind Control tasks and two Atari environments using DreamerV3 as the underlying learner as well as twelve state-based DeepMind Control tasks using MBPO as the backbone. The learning curves for vision-based tasks are given in 10 and for state-based tasks in Figures 11. Across these tasks, VDO-grad and VDO-spread remain broadly competitive with SOMBRL and improve over the unmodified backbone on tasks where exploration is important such as Hopper Hop and Enduro.

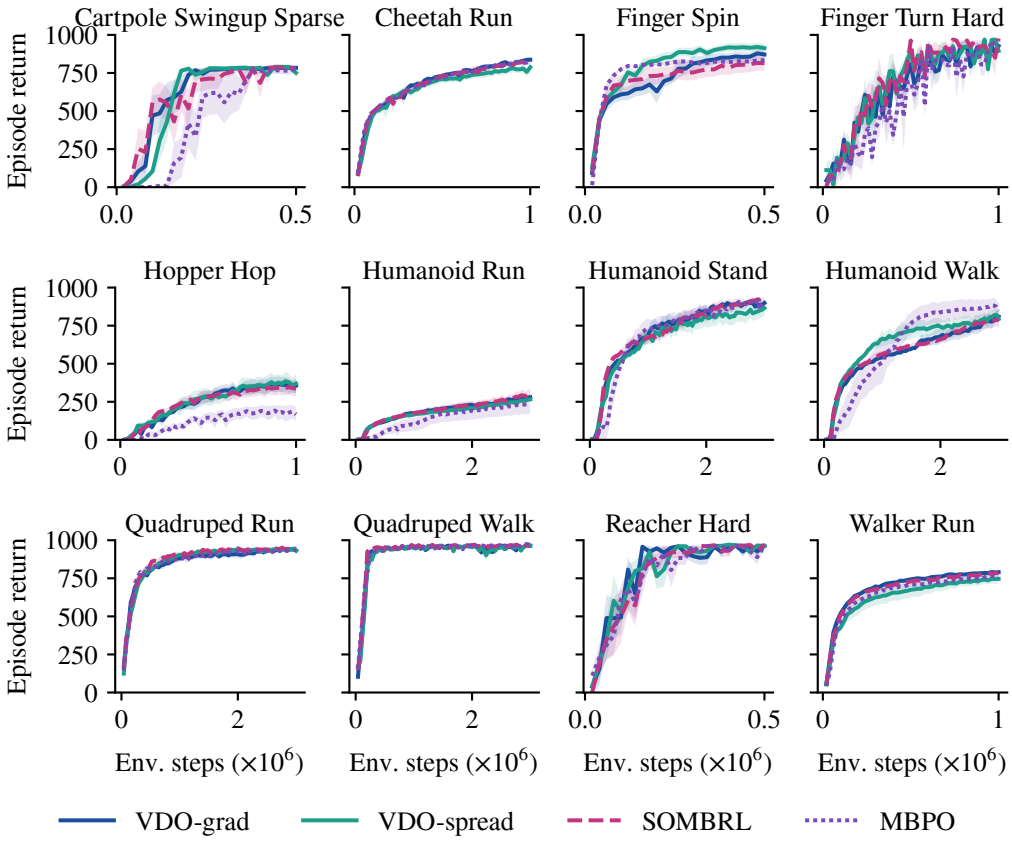


Figure 11: Learning curves over 5 seeds on state-based DMC tasks with MBPO as the base learner.