Information-Based Exploration via Random Features for Reinforcement Learning

Waris Radji WARIS.RADJI@INRIA.FR
Odalric-Ambrym Maillard ODALRIC.MAILLARD@INRIA.FR
Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189-CRIStAL, F-59000 Lille, France

Editors: Hung-yi Lee and Tongliang Liu

Abstract

Representation learning has enabled classical exploration strategies to be extended to deep Reinforcement Learning (RL), but often makes algorithms more complex and theoretical guarantees harder to establish. We introduce Random Feature Information Gain (`RFIG`), grounded in Bayesian kernel methods theory, which uses random Fourier features to approximate information gain and compute exploration bonuses in non-countable spaces. We provide error bounds on information gain approximation and avoid the black-box aspects of neural network-based uncertainty estimation, for optimism-based exploration. We present practical details that make `RFIG` scalable to deep RL scenarios, enabling smooth integration into standard deep RL algorithms. Experimental evaluation across diverse control and navigation tasks demonstrates that `RFIG` achieves competitive performance with well-established deep exploration methods while offering superior theoretical interpretation.

Keywords: Reinforcement Learning; Exploration; Random Features; Kernel Methods;

1. Introduction

In Reinforcement Learning (RL), agents learn optimal decision-making strategies through trial-and-error interactions with an environment, receiving rewards or penalties that guide their learning process (Sutton et al., 1998). A fundamental challenge is the exploration-exploitation tradeoff, where agents must balance between exploiting current knowledge to maximize immediate rewards and exploring new actions to potentially discover better long-term strategies. In this paper, we focus our attention on the exploration part in continuous and high-dimensional problems. In small-scale environments like Multi-Armed Bandits (MABs) and discrete Markov Decision Processes (MDPs), an effective strategy is the Optimism in Face of Uncertainty (OFU), which operates on the principle that when an agent lacks sufficient information about certain states, it should assume they may yield high rewards, thereby encouraging exploration of these uncertain regions (Auer et al., 2002, 2008). This principle, theoretically grounded in MABs, was adapted to more general problems, through an *exploration bonus*, where the reward obtained by the learner is augmented, typically $r_t + \beta r_t^+$, where at interaction step t, r_t is the reward given by the MDP, r_t^+ is the exploration bonus and $\beta \geq 0$ is an parameter that control exploration strength. The literature often considers bonus in the form of $1/\sqrt{n_t(s)}$, where $n_t(s)$ is the number of times the agent has visited state s at interaction t: the more we visit a state, the more certain we are about the estimations (Strehl and Littman, 2008). However, in *non-countable* spaces, which

© 2025 W. Radji & O.-A. Maillard.

we consider here, this bonus is not straightforward to implement: the probability of visiting the same state twice can be zero, and direct count-based exploration becomes meaningless.

Research question. This fundamental challenge has led to the development of deep learning-based exploration strategies, where neural networks (NNs) exploit *representation learning* to learn a proxy of uncertainty or *pseudo-counts*. Traditional deep representation learning approaches, while empirically successful (Bellemare et al., 2016; Pathak et al., 2017; Badia et al., 2020), suffer from limited interpretability that hinders theoretical understanding and creates hyperparameter sensitivity. This brittleness arises from complex interactions between optimization dynamics and problem structure, where gradient-based algorithms and backpropagation show fragile dependencies on learning rate schedules that can destabilize training and require parameter adjustments spanning orders of magnitude across domains (Glorot and Bengio, 2010). Additionally, architectural choices and regularization strategies depend on environmental characteristics, creating optimization landscapes where effective hyperparameter settings often transfer poorly and fail dramatically in new contexts. This sensitivity means that even algorithmically sound approaches can exhibit dramatic performance degradation when operating under suboptimal hyperparameter regimes (Henderson et al., 2018). These observations raise an interesting question:

How can we design alternative exploration mechanisms for deep RL that offer theoretical interpretability and computational tractability while achieving competitive empirical performance relative to standard deep methods?

Interesting directions. A promising direction lies in kernel methods, which provide theoretically grounded uncertainty quantification through closed-form solutions rather than iterative gradient-based optimization. These methods avoid the extensive hyperparameter tuning and training instabilities inherent to neural architectures (Srinivas et al., 2009), while maintaining the ability to capture complex nonlinear patterns. From a *Bayesian* perspective, the concept of *information gain*, provides a grounded approach to create exploration bonus that go beyond discrete spaces: poorly visited states are very uncertain and could lead to high information gain, making them attractive targets for exploration while naturally diminishing the bonus as states become well-explored and their uncertainty decreases (Kolter and Ng, 2009). However, *vanilla* kernel methods suffer from cubic computational complexity, limiting their scalability to the large sample sizes required in deep RL.

Contributions. In this paper, to tackle the problem of OFU exploration in uncountable spaces, we introduce a novel exploration bonus for deep RL: Random Feature Information Gain (`RFIG`). Our bonus is directly derived from information gain quantification in Bayesian kernel methods, which we combine with recent advances that allow these methods to become scalable. An important component of our approach is to exploit random features (Rahimi and Recht, 2007), which enable capturing complex nonlinear spatial patterns in high-dimensional data and approximate kernels, to handle the cubic scaling in the number of samples. Unlike pure deep learning approaches, `RFIG` computes exploration bonuses via closed-form solutions, eliminating backpropagation complexity and hyperparameter brittleness while maintaining theoretical interpretability.

Outline. We first derive `RFIG` from Bayesian kernel methods and random features (Section 4.1). We then provide approximation error bounds (Section 4.2) and apply them to

random Fourier features (Section 4.3). Finally, we detail algorithmic integration with deep RL (Section 5.1) and demonstrate effectiveness across diverse tasks (Section 5.2).

2. Related Work

Exploration remains a fundamental challenge in RL, particularly in environments with sparse rewards or large state spaces. We review existing approaches, progressing from general methods through kernelized MDPs to deep RL exploration strategies.

Exploration foundations. The exploration-exploitation tradeoff was first formalized in MABs through OFU-based algorithms (Auer et al., 2002) and Thompson Sampling (Thompson, 1933; Chapelle and Li, 2011). Recent approaches like Information Directed Sampling (Russo and Van Roy, 2014) and Minimum Empirical Divergence (Honda and Takemura, 2010) directly formalize information gain in their objectives. These principles have been extended to tabular MDPs (Auer et al., 2008; Osband et al., 2013; Pesquerel and Maillard, 2022) and subsequently to continuous spaces through kernelization and deep RL.

Kernelized MDPs. Kernel methods extend bandit exploration principles to MDPs with theoretically grounded uncertainty quantification. In bandits, `GP-UCB` achieves provable regret guarantees using Gaussian Process posteriors (Srinivas et al., 2009; Valko et al., 2013). This framework extends to MDPs where kernels encode similarity structure over state-action spaces (Morere and Ramos, 2018; Chowdhury and Gopalan, 2019; Domingues et al., 2021), and demonstrates that kernel-based uncertainty quantification enables theorically-grounded exploration in continuous MDPs but remains limited to relatively small-scale problems due to computational constraints.

Deep RL exploration. Scaling to high-dimensional spaces typically requires representation learning. Curiosity-driven methods (Pathak et al., 2017) and episodic novelty approaches (Badia et al., 2020) learn embeddings for exploration. Information-theoretic methods maximize information gain through NN ensembles (Houthooft et al., 2016; Nikolov et al., 2018; Sukhija et al., 2024) or prediction disagreement (Osband et al., 2016; Azizzadenesheli et al., 2018). While effective, these approaches couple exploration with representation learning, creating hyperparameter sensitivity (Glorot and Bengio, 2010; Henderson et al., 2018). A smaller line of work separates exploration from representation learning. `RND` (Burda et al., 2018) uses random features by training a network to predict a fixed random network's outputs, where prediction error signals novelty. This demonstrates feature learning is not required for effective exploration, though `RND` lacks theoretical grounding. Yang et al. (2024) addresses this through connections to pseudo-counts.

Bridging kernels and deep RL. Recent attempts to apply kernel methods in deep RL reveal different approaches. Ma et al. (2024) uses random Fourier features with kernel density estimation but requires user-defined success criteria for specific environments. Blau et al. (2019) develops a "Bayesian curiosity module" using posterior variance from learned kernels but suffers from cubic complexity, limiting scalability. They mention RFFs as future work, which we implement in this paper.

3. Background on Information Gain, RL and Scalable Kernels

This section establishes the theoretical foundations: information gain for exploration, Bayesian kernel methods for uncertainty quantification, and random Fourier features for scalability.

Exploration in RL via information gain. An agent interacts with a discounted MDP $\mathbf{M} = (\mathcal{S}, \mathcal{A}, \mathbf{r}, \mathbf{p}, \gamma)$ to learn a policy $\pi : \mathcal{S} \to \Pr(\mathcal{A})$ maximizing expected cumulative reward $J(\pi) = \mathbb{E}_{\pi,\mathbf{p}}\left[\sum_{t=0}^{\infty} \gamma^t \mathbf{r}(s_t, a_t)\right]$ (Sutton et al., 1998). A standard approach augments rewards with exploration bonuses (Strehl and Littman, 2008): $\mathbf{r}_{\text{total}}(s,a) = \mathbf{r}(s,a) + \beta \mathbf{r}^{+}(s,a)$, where $\beta > 0$ controls exploration strength. The widely-used bonus $1/\sqrt{n(s)}$ (with $n(s)$ the visit count for state s) implicitly maximizes information gain (Bellemare et al., 2016). To formalize this, consider learning an unknown function $f : \mathcal{X} \to \mathbb{R}$ from noisy data $\mathcal{D}_n = \{(x_i, y_i)\}_{i=1}^n$ where $y_i = f(x_i) + \eta_i$. The Bayesian posterior $p(f \mid \mathcal{D}_n)$ encodes uncertainty about f. The expected information gain from querying x_* is

$$\text{IG}(x_* \mid \mathcal{D}_n) = H(f \mid \mathcal{D}_n) - \mathbb{E}_{Y_*}[H(f \mid \mathcal{D}_n \cup \{(x_*, Y_*)\})] \tag{1}$$

where $H(f \mid \mathcal{D}) = -\int p(f \mid \mathcal{D}) \log p(f \mid \mathcal{D})\, df$ is the differential entropy (Cover, 1999). This criterion, connect with active inference (Settles, 2009; Friston et al., 2015).

Bayesian kernel methods. As a alternative to neural exploration, we employ kernel methods that provide principled uncertainty quantification through implicit mapping to reproducing kernel Hilbert spaces $\mathcal{H}_k$ (Aronszajn, 1950; Schölkopf et al., 2001). A positive semi-definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ enables high-dimensional computations using only pairwise similarities. In Bayesian kernel ridge regression (Saunders et al., 1998; Jaakkola and Haussler, 1999) with regularization $\lambda > 0$, the posterior mean and variance are

$$\mu_n(x) = \mathbf{k}_n(x)^T(\mathbf{K}_n + \lambda \mathbf{I}_n)^{-1}\mathbf{y}_n \qquad \sigma_n^2(x) = k(x,x) - \mathbf{k}_n(x)^T(\mathbf{K}_n + \lambda \mathbf{I}_n)^{-1}\mathbf{k}_n(x) \tag{2}$$

where $\mathbf{K}_n \in \mathbb{R}^{n \times n}$ with $[\mathbf{K}_n]_{ij} = k(x_i, x_j)$ and $\mathbf{k}_n(x) = [k(x_1, x), \ldots, k(x_n, x)]^T$. This is equivalent to a Gaussian process $f \sim \mathcal{GP}(0, k(x, x'))$ with noise $\sigma^2 = \lambda$ (Williams and Rasmussen, 1995). However, inverting $(\mathbf{K}_n + \lambda \mathbf{I}_n)$ requires $\mathcal{O}(n^3)$ operations, prohibitive for large datasets. Random Fourier Features (RFFs) (Rahimi and Recht, 2007) resolve this bottleneck by approximating kernels with explicit finite-dimensional mappings. For shift-invariant kernels $k(x, x') = k(x - x')$, Bochner's theorem (Bochner et al., 1959) enables $k(x, x') \approx \boldsymbol{\phi}(x)^T \boldsymbol{\phi}(x')$ where

$$\boldsymbol{\phi}(x) = \sqrt{\frac{2}{D}} \begin{bmatrix} \cos(\boldsymbol{\omega}_1^T x + b_1) \\ \vdots \\ \cos(\boldsymbol{\omega}_D^T x + b_D) \end{bmatrix} \tag{3}$$

with $\boldsymbol{\omega}_i \sim p(\boldsymbol{\omega})$ from the kernel's spectral density (Fourier transform of k) and $b_i \sim \text{Uniform}[0, 2\pi]$. For the RBF kernel $k(x, x') = \exp(-\|x - x'\|^2 / 2\ell^2)$, the length-scale ℓ determines $p(\boldsymbol{\omega}) = \mathcal{N}(0, \ell^{-2}\mathbf{I})$. Applying the Woodbury identity (Woodbury, 1950) with feature matrix $\boldsymbol{\Phi}_n \in \mathbb{R}^{n \times D}$ yields

$$\mu_n(x) = \boldsymbol{\phi}(x)^T(\boldsymbol{\Phi}_n^T \boldsymbol{\Phi}_n + \lambda \mathbf{I}_D)^{-1} \boldsymbol{\Phi}_n^T \mathbf{y}_n \tag{4}$$

$$\sigma_n^2(x) = \boldsymbol{\phi}(x)^T \boldsymbol{\phi}(x) - \boldsymbol{\phi}(x)^T(\boldsymbol{\Phi}_n^T \boldsymbol{\Phi}_n + \lambda \mathbf{I}_D)^{-1} \boldsymbol{\phi}(x). \tag{5}$$

This reduces computational complexity from $\mathcal{O}(n^3)$ to $\mathcal{O}(D^3)$, enabling efficient uncertainty quantification that scales with feature dimension D rather than dataset size n.

4. Random Feature Information Gain

Before looking at how information gain is implemented in a RL training loop to promote exploration, we now derive our Random Feature Information Gain (`RFIG`). The derivation proceeds in three steps: (1) express GP information gain in terms of posterior variance, (2) approximate the kernel matrix using random features, (3) apply matrix identities to obtain the final $\mathcal{O}(D^3)$ form. All detailed proofs of this section can be found in Appendix A.

4.1. Derivation

We start by recalling the information gain in the Gaussian process framework using the entropy reduction formulation, as described in (1). Consider a Gaussian process, that we defined in Section 3, $f \sim \mathcal{GP}(0, k(\cdot,\cdot))$ with observation noise $\eta \sim \mathcal{N}(0, \sigma^2)$. Given current data $\mathcal{D}_n = \{(x_i, y_i)\}_{i=1}^n$, the posterior entropy can be expressed using the kernel matrix $\mathbf{K}_n$ with $H(f \mid \mathcal{D}_n) = \frac{1}{2}\log\det(2\pi e(\mathbf{K}_n + \sigma^2\mathbf{I}_n)^{-1})$. When we add a new observation (x_*, y_*) to our dataset, obtaining $\mathcal{D}_{n+1} = \mathcal{D}_n \cup \{(x_*, y_*)\}$, the posterior distribution changes.

Definition 1 (Information gain in GP (Lawrence et al., 2002)) [1] *The information gain, as defined in (1), can be expressed for a query point x_* in GP, as*

$$\begin{aligned}\mathrm{IG}(x_* \mid \mathcal{D}_n) &= H(f \mid \mathcal{D}_n) - \mathbb{E}_{Y_*}[H(f \mid \mathcal{D}_{n+1})] \\ &= \frac{1}{2}\log\det(2\pi e(\mathbf{K}_n + \sigma^2\mathbf{I}_n)^{-1}) - \mathbb{E}_{Y_*}\left[\frac{1}{2}\log\det(2\pi e(\mathbf{K}_{n+1} + \sigma^2\mathbf{I}_{n+1})^{-1})\right] \\ &= \frac{1}{2}\log\det(\mathbf{K}_{n+1} + \sigma^2\mathbf{I}_{n+1}) - \frac{1}{2}\log\det(\mathbf{K}_n + \sigma^2\mathbf{I}_n) = \boxed{\frac{1}{2}\log\left(1 + \frac{\sigma_n^2(x_*)}{\sigma^2}\right)}.\end{aligned}$$

Challenges. However, computing $\sigma_n^2(x_*)$ requires inverting the $n \times n$ matrix $(\mathbf{K}_n + \sigma^2\mathbf{I}_n)$, which scales as $\mathcal{O}(n^3)$ and becomes prohibitive for huge datasets. To address this computational bottleneck, we next develop a random feature approximation that reduces complexity from $\mathcal{O}(n^3)$ to $\mathcal{O}(D^3)$, where D is the number of features (Proposition 2). Additionally, practical implementation requires careful selection of the kernel length-scale ℓ, which controls the smoothness assumptions and generalization behavior, and the number of random features D, which determines the approximation quality versus computational cost trade-off. Later in the paper, we will propose insights for choosing D based on theoretical error bounds (Corollary 9) and adaptive length-scale selection recommandations (Section 5.1).

Proposition 2 (Information Gain via Random Features) *Consider a random feature transformation $\boldsymbol{\phi} : \mathcal{X} \to \mathbb{R}^D$ that approximates a shift-invariant kernel $k(x, x') \approx \boldsymbol{\phi}(x)^T\boldsymbol{\phi}(x')$ (Rahimi and Recht, 2007). The information gain (Definition 1) can be approximated as*

$$\hat{\mathrm{IG}}(x_* \mid \mathcal{D}_n) = \frac{1}{2}\log\left(1 + \boldsymbol{\phi}(x_*)^T(\boldsymbol{\Phi}_n^T\boldsymbol{\Phi}_n + \lambda\mathbf{I}_D)^{-1}\boldsymbol{\phi}(x_*)\right) \tag{6}$$

where $\boldsymbol{\Phi}_n \in \mathbb{R}^{n\times D}$ is the feature matrix with rows $\boldsymbol{\phi}(x_i)^T$ for $i = 1, \ldots, n$, and $\lambda > 0$ is the regularization parameter.

1. This formulation is equivalent to what they term the "differential entropy score".

4.2. Error Bounds

To provide theoretical guarantees for our approach, we establish error bounds for RFIG under uniform kernel convergence assumptions. Our analysis serves two key purposes: (1) quantifying how errors in kernel approximation propagate to information gain estimates, and (2) determining the number of random features D required to achieve a desired approximation accuracy ε with high probability. We proceed by first bounding the error in posterior variance estimation, then using this result to establish guarantees for information gain approximation, and finally applying our general framework to RFFs. Our analysis relies on three standard assumptions commonly employed in the random features literature (Rahimi and Recht, 2007; Sutherland and Schneider, 2015).

Assumption 3 (Uniform kernel approximation) *The random feature map $\boldsymbol{\phi}(x) : \mathcal{X} \to \mathbb{R}^D$ provides a uniform approximation to the kernel $k(x, x')$ over the domain:*

$$\mathbb{P}\Big[\sup_{x,x' \in \mathcal{X}} |\boldsymbol{\phi}(x)^\top \boldsymbol{\phi}(x') - k(x, x')| \geq \epsilon\Big] \leq \delta(\epsilon; d, D). \tag{7}$$

Assumption 4 (Regularization scaling) *The regularization parameter scales linearly with sample size: $\lambda = n\lambda_0$ for some $\lambda_0 > 0$.*

Assumption 5 (Bounded kernel) *The kernel is bounded: $|k(x, x')| \leq \kappa$ for all $x, x' \in \mathcal{X}$.*

Assumption 3 is the core requirement for random feature methods and holds for RFFs under mild conditions on the input domain (Rahimi and Recht, 2007). Assumption 4 ensures that the regularization term remain properly balanced as sample size grows, preventing regularization from either dominating or vanishing asymptotically, which is useful for deriving clean convergence rates and consistency results. Assumption 5 is satisfied by most practical kernels including RBF and Matérn kernels.

Posterior variance error. Since information gain is fundamentally determined by posterior variance (1), we first establish how kernel errors propagate to variance estimates.

Proposition 6 (Posterior variance error bound) *Under Assumptions 3, 4, and 5, the error in posterior variance estimation when using random features is bounded by:*

$$\forall x \in \mathcal{X}, |\hat{\sigma}_n^2(x) - \sigma_n^2(x)| \leq \epsilon \left(1 + \frac{\kappa^2}{\lambda_0^2} + \frac{2\kappa}{\lambda_0} + \frac{\epsilon}{\lambda_0}\right), \tag{8}$$

where $\epsilon = \sup_{x,x' \in \mathcal{X}} |\boldsymbol{\phi}(x)^\top \boldsymbol{\phi}(x') - k(x, x')|$.

This result shows that variance estimation error scales linearly with kernel approximation quality ϵ and exhibits the expected dependence on regularization strength.

Information gain error. The connection between posterior variance and information gain enables us to translate variance errors into information gain guarantees (Lemma 10).

Proposition 7 (RFIG error bound) *Under Assumptions 3, 4, and 5, the error in RFIG approximation is bounded by:*

$$|\operatorname{IG}(x|\mathcal{D}_n) - \hat{\operatorname{IG}}(x|\mathcal{D}_n)| \leq \frac{\epsilon(\lambda_0+\kappa)^2 + \epsilon^2\lambda_0}{2n\lambda_0^3}, \tag{9}$$

where $\epsilon = \sup_{x,x'\in\mathcal{X}} |\boldsymbol{\phi}(x)^\top \boldsymbol{\phi}(x') - k(x,x')|$.

Our bound exhibits some properties: the error decreases with sample size n (consistency), scales with kernel approximation quality ϵ (approximation dependence), and reveals a regularization trade-off where stronger λ_0 tightens the bound but may over-smooth posteriors.

4.3. Application to Random Fourier Features

We apply our general bound to RFFs by using existing uniform convergence results.

Proposition 8 (RFF uniform convergence Rahimi and Recht (2007)) *Let* $\mathcal{X} \subset \mathbb{R}^d$ *be compact with diameter* $\operatorname{diam}(\mathcal{X})$ *and* k *a shift-invariant kernel with unit maximum and Fourier transform* $P(\omega)$*. Let* $\sigma_p^2 = \mathbb{E}_P[\|\omega\|^2]$*. For RFF mapping* $\boldsymbol{\phi}$ *and any* $\epsilon > 0$*:*

$$\Pr\left[\|\boldsymbol{\phi}^\top\boldsymbol{\phi} - k\|_\infty \geq \epsilon\right] \leq c\left(\frac{\sigma_p \operatorname{diam}(\mathcal{X})}{\epsilon}\right)^2 \exp\left(-\frac{D\epsilon^2}{8(d+2)}\right), \tag{10}$$

$c = 256$ *in Rahimi and Recht (2007), tightened to 66 in Sutherland and Schneider (2015).*

Corollary 9 (Feature dimension requirement) *To achieve approximation error* $\sup_x |\operatorname{IG}(x|\mathcal{D}_n) - \hat{\operatorname{IG}}(x|\mathcal{D}_n)| \leq \varepsilon$ *with probability at least* $1-\delta$*, it suffices to choose*

$$D = \mathcal{O}\left(\frac{d}{\epsilon_k^2}\log\frac{\sigma_p \operatorname{diam}(\mathcal{X})}{\epsilon_k\delta}\right), \tag{11}$$

where $\epsilon_k = \dfrac{2n\lambda_0^3\varepsilon}{(\lambda_0+\kappa)^2}$ *when* ε *is sufficiently small.*

Even if $\operatorname{diam}(\mathcal{X})$ is generally not know in a RL context, this result provides practical guidance for hyperparameter selection: the required feature dimension D scales linearly with problem dimension d and logarithmically with desired accuracy. Importantly, D decreases with sample size n through ϵ_k, reflecting that larger datasets permit coarser kernel approximations while maintaining the same information gain accuracy. This theoretical foundation justifies our approach and enables confident deployment in practical exploration scenarios.

5. `RFIG` for Efficient Exploration in RL

This paper aims to apply `RFIG` for improving optimism-based exploration in deep RL. This section outlines the key algorithmic components and implementation considerations that enable efficient and scalable integration with existing deep RL agents.

5.1. Details that Matter

While Algorithm 1 outlines the core `RFIG` integration with deep RL, successful implementation requires attention to several practical considerations. However, these hyperparameter choices are minimal compared to neural network approaches, which typically demand extensive tuning of learning rates, network architectures, regularization schemes, and optimization schedules. This subsection presents the key considerations that determine `RFIG`'s effectiveness in practice, demonstrating the relative simplicity of our kernel-based approach.

Algorithm 1: `RFIG` for exploration

Input: RFF Feature map $\boldsymbol{\phi}_\ell : \mathcal{X} \to \mathbb{R}^D$ with length-scale $\ell \propto \sqrt{\bar{d}}$, regularization $\lambda > 0$, subsample ratio $\rho \in (0, 1]$, environment $\mathbf{M}$, policy π, exploration scale $\beta > 0$.

Initialize RFIG matrices $\boldsymbol{\Sigma}_0 \leftarrow \lambda \mathbf{I}_D$ and $\boldsymbol{\Lambda}_0 \leftarrow \lambda^{-1}\mathbf{I}_D$

Initialize state normalization parameters (μ_s, σ_s^2)

for $t \leftarrow 1, 2, \cdots$ **do**

- Collect N transitions $\mathcal{D} = \{(s_i, a_i, r_i, s_i')\}_{i=1}^N$ with policy π in environment $\mathbf{M}$
- Update normalization parameters with $\{s_i\}_{i=1}^N$, obtain normalized states $\{\bar{s}_i\}_{i=1}^N$
- Compute information gain bonuses $\mathcal{R}^+ = \left\{r_i^+ = \frac{1}{2}\log\left(1 + \boldsymbol{\phi}_\ell(\bar{s}_i)^\top \boldsymbol{\Lambda}_{t-1}\boldsymbol{\phi}_\ell(\bar{s}_i)\right)\right\}_{i=1}^N$
- Subsample $\lfloor N\rho \rfloor$ states uniformly from $\{\bar{s}_i\}_{i=1}^N$ to form $\boldsymbol{\Phi}_t$ with rows $\boldsymbol{\phi}_\ell(\bar{s}_j)^\top$
- Update $\boldsymbol{\Sigma}_t \leftarrow \boldsymbol{\Sigma}_{t-1} + \boldsymbol{\Phi}_t^\top \boldsymbol{\Phi}_t$, then $\boldsymbol{\Lambda}_t \leftarrow \boldsymbol{\Sigma}_t^{-1}$
- Update policy π using any RL algorithm with $\mathcal{D}$ and bonuses $\mathcal{R}^+$

end

length-scale selection. The length-scale ℓ controls the smoothness of the uncertainty estimates and should account for the curse of dimensionality. In high-dimensional spaces, typical distances between points scale as $\sqrt{\bar{d}}$ where $\bar{d}$ is the *effective* input dimension (Hvarfner et al., 2024; Xu et al., 2024). Therefore, we recommend initializing $\ell \propto \sqrt{\bar{d}}$. To estimate the effective dimension from samples, we refer to Section 4 of Valko et al. (2013).

State normalization[2]. We maintain running statistics μ_s and σ_s^2 to normalize states as $\bar{s} = (s - \mu_s)/\sigma_s$. This prevents scale differences across dimensions from dominating kernel computations and is critical for RFF effectiveness.

Subsampling Strategy[2]. The subsample ratio ρ serves multiple purposes. The primary goal is to prevent information gain from shrinking too rapidly to zero as the number of samples grows, which would lead to premature exploration termination. Additionally, subsampling helps Newton-Schulz iterations converge faster since the covariance matrix $\boldsymbol{\Sigma}_t$ changes more slowly between updates, making warm starts more effective. This approach mirrors techniques in sparse Gaussian processes, where a subset of inducing points can represent the uncertainty structure of the entire dataset.

Newton-Schulz matrix inversion. A key computational challenge in RFIG is efficiently maintaining the matrix $(\boldsymbol{\Phi}_n^T \boldsymbol{\Phi}_n + \lambda \mathbf{I}_D)^{-1}$ as new observations arrive. It's possible to employ the Newton-Schulz iteration (Schulz, 1933), which iteratively computes matrix inverses using

2. These details have shown beneficial for many deep exploration strategies in Yuan et al. (2024).

$\mathbf{X}_{k+1} = \mathbf{X}_k(2\mathbf{I} - \mathbf{A}\mathbf{X}_k)$. This method converges quadratically to $\mathbf{A}^{-1}$ when $\|\mathbf{I} - \mathbf{A}\mathbf{X}_0\|_2 < 1$ and crucially allows using the previous iteration's result as a warm start for $\mathbf{X}_0$. Compared to Sherman-Morrison or Woodbury updates, more commonly considered, Newton-Schulz offers superior numerical stability by avoiding explicit small-number divisions and provides computational savings. Combined with its parallel structure that maps naturally to GPU architectures, Newton-Schulz is ideally suited for the frequent matrix updates required in online deep RL applications. Further details are in Appendix B.1.

5.2. Numerical Experiments

We evaluate `RFIG` by integrating it with Proximal Policy Optimization (`PPO`) algorithm (Schulman et al., 2017), following the non-episodic exploration framework described in Burda et al. (2018) for Random Network Distillation (`RND`). This allows for direct comparison with proven implementation practices for intrinsic motivation in deep RL. Following the `PPO+RND` architecture, we augment the standard `PPO` objective with `RFIG`-based intrinsic rewards. We maintain separate value networks for extrinsic and intrinsic rewards and normalize both rewards, as this has benefited many bonuses (Yuan et al., 2024).

Baselines. In addition to `RND`, we consider two other deep RL baselines that also follow the optimism under the face of uncertainty (OFU) principle. We include `#Explo` from Tang et al. (2017), which learns a hash function through an autoencoder architecture, maintains a hash table with visit counts, and employs standard count-based exploration bonuses. We also consider `VIME` (Houthooft et al., 2016), which shares a similar spirit to our approach by targeting information gain for exploration. `VIME` learns dynamics with a Bayesian NN, a NN with probability distributions over weights rather than fixed parameters, and approximates information gain as the Kullback-Leibler divergence between the prior and posterior weight distributions of the network. Both methods require effective representation learning to function properly and reflect well the popular neural network-based exploration paradigm that works independently from the policy learning process, via reward bonuses.

Setup. We adopt global hyperparameter settings proven effective for `PPO` across all experiments (detailed in Appendix B.2). The exploration coefficient β is set to 0.5 for all methods. Since we normalize exploration bonuses before integration, this coefficient does not affect the relative comparison between baselines. For fair comparison across exploration baselines, we avoid extensive hyperparameter search and instead use common, well-established parameter values found in reference implementations. All baseline methods employ neural networks with 256×256 hidden layers and perform one gradient step per batch update using the Adam optimizer. We initialize observation normalization with random trajectories for all methods and additionally estimate the effective dimension $\bar{d}$ for `RFIG`. For `RFIG`-specific parameters, we use $D = 1024$ random features, regularization $\lambda = 1$, subsample ratio $\rho = 6.25\%$, and length-scale $\ell = \sqrt{\bar{d}}$. We evaluate `RFIG` across four domains designed to test exploration capabilities. Classic control tasks (Acrobot, MountainCar) provide baseline comparisons in low-dimensional settings (Lange, 2022). For challenging continuous control, we use sparse reward variants of Brax locomotion tasks (Freeman et al., 2021), where agents receive milestone rewards only upon reaching specific distance thresholds (Appendix B.3). We include PointMaze navigation environments (de Lazcano et al., 2024; Radji, 2025) and

MinAtar tasks (Young and Tian, 2019; Lange, 2022), miniature Atari implementations that demonstrate exploration needs extend beyond sparse reward settings. All experiments use 32 random seeds and 32 parallel environments with 128-step unrolls.

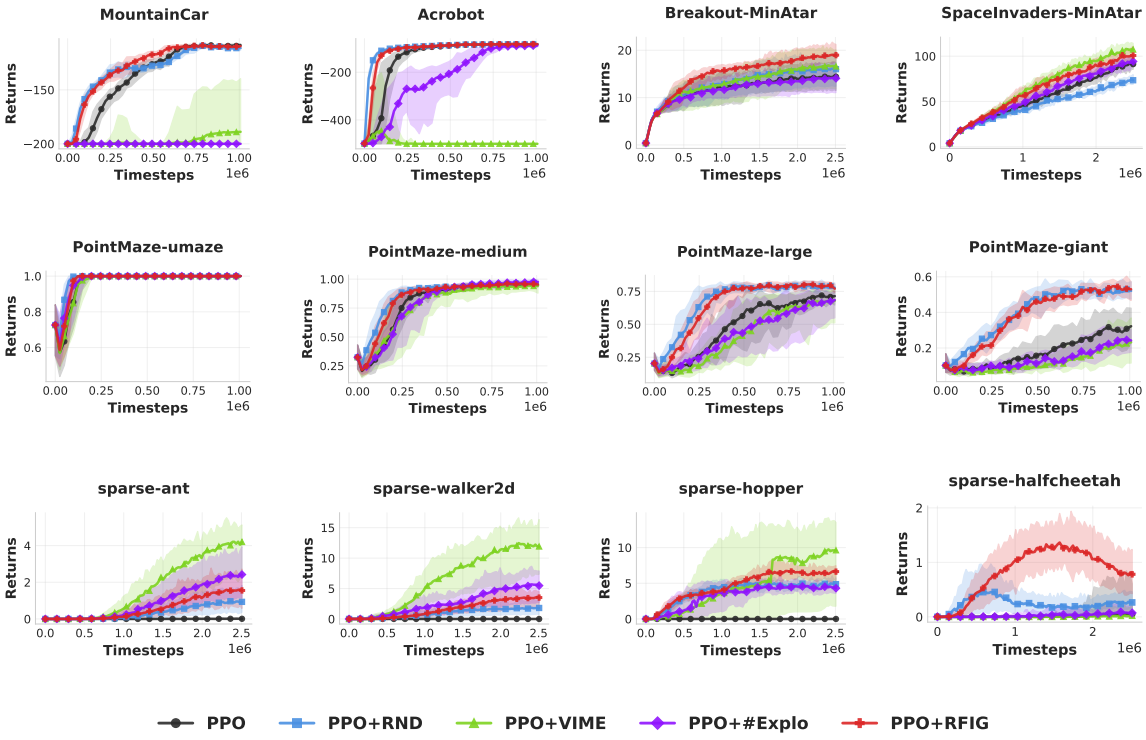

Figure 1: Comparing exploration strategies. Solid lines represent the interquartile mean, with shaded areas indicating the 25th-75th percentiles across 32 random seeds.

Discussion. The experimental results across tasks of varying dimensionality reveal distinct performance patterns that highlight fundamental tradeoffs in exploration strategy design (Table 1). `PPO+RFIG` demonstrates remarkable consistency, outperforming vanilla `PPO`, which demonstrates that our exploration bonus is effective. The key question is how well it performs relative to established methods. `RFIG` is competitive with NN-based methods across all environments. From control tasks to complex navigation challenges, `RFIG` maintains stable performance without catastrophic failures. The representation-dependent methods exhibit environment-specific behavior. `VIME` excels in some high-dimensional continuous control tasks like sparse-ant and sparse-walker2d when its Bayesian NN representations seem to succeed, but struggles in some other tasks, which can be a failure in capturing a good representation. `#Explo` shows similar inconsistency. We believe that both approaches could likely improve with environment-specific hyperparameter tuning, which undermines their general direct applicability. In contrast, `RFIG` avoids this representational brittleness. The comparison with `RND` proves particularly revealing since both employ random features through different mechanisms. Beyond empirical performance, `RFIG` offers practical ad-

vantages through its closed-form solution approach. Our JAX implementation[3] shows no significant execution time differences compared to NN approaches, even with full matrix inversion used in our final experiments. `RFIG` competitively matches or exceeds `RND`'s performance while offering superior theoretical interpretability. Moreover, `RFIG` achieves peak performance in challenging environments like sparse-halfcheetah and Breakout, demonstrating effectiveness across diverse reward structures.

Method	Principle	Representation	Computation	Sensitivity
RND	Prediction error	✗	Distillation	Medium
VIME	Information gain	✓	Bayesian NN	High
#Explo	Count-based	✓	Auto-encoder	High
RFIG	**Information gain (Kernel theory)**	✗	**Closed-form**	**Low**

Table 1: Comparison of exploration methods across key characteristics.

6. Conclusion

We introduced the Random Feature Information Gain (`RFIG`) exploration bonus, demonstrating that principled kernel methods can match the empirical performance of neural network-based exploration while offering some theoretical insights and closed-form solutions. By leveraging Bayesian kernel methods and random Fourier features, `RFIG` achieves competitive results across diverse domains without requiring representation learning or hyperparameter tuning. Our work challenges the prevailing assumption that effective exploration requires increasingly sophisticated neural architectures, which often have a black-box aspect. `RFIG`'s success stems from its theoretical foundation in kernel methods, providing mathematical rigor often absent in popular deep RL exploration strategies

Broader impact. The framework underlying `RFIG` and its theoretical guarantees extends naturally to other domains requiring uncertainty quantification. Active learning, Bayesian optimization, out-of-distribution detection in offline RL, and other sequential decision-making problems can all benefit from similar kernel-based approaches.

Future directions. Several promising research avenues emerge from this work. First, studying how `RFIG` behaves in very high-dimensional spaces like big images will reveal the scalability limits of our kernel-based approach and identify potential adaptations needed for visual domains. Second, developing adaptive mechanisms for kernel parameter selection, particularly length-scale tuning, to potentially improve performance. Third, while we focused on RBF kernels and their RFFs approximation, exploring alternative kernels through different random projection schemes offers exciting possibilities. Specialized kernels might close the performance gap with deep RL methods that currently outperform RFIG in certain environments. Finally, the same information-theoretic principles could enable conservative exploration strategies for offline RL, where avoiding out-of-distribution states takes precedence over optimistic exploration.

3. The experiment code is available at https://github.com/riiswa/rfig.

Acknowledgments

We thank Emilie Kaufmann for valuable feedback on this work. The authors are affiliated with the Inria Scool team project. This work has been supported by the French Ministry of Higher Education and Research, the Hauts-de-France region, Inria, and the MEL. Additional support was provided by the French National Research Agency under the PEPR IA FOUNDRY project (ANR-23-PEIA-0003) and the ANR JCJC REPUBLIC project (ANR-22-CE23-0003-01). Experiments presented in this paper were carried out using the PlaFRIM experimental testbed, supported by Inria, CNRS (LABRI and IMB), Université de Bordeaux, Bordeaux INP and Conseil Régional d'Aquitaine (see https://www.plafrim.fr).

References

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

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

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

Kamyar Azizzadenesheli, Emma Brunskill, and Animashree Anandkumar. Efficient exploration through bayesian deep q-networks. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–9. IEEE, 2018.

Adrià Puigdomènech Badia, Pablo Sprechmann, Alex Vitvitskyi, Daniel Guo, Bilal Piot, Steven Kapturowski, Olivier Tieleman, Martín Arjovsky, Alexander Pritzel, Andew Bolt, et al. Never give up: Learning directed exploration strategies. *arXiv preprint arXiv:2002.06038*, 2020.

Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and hashing. In *Advances in neural information processing systems*, pages 2611–2619, 2016.

Tom Blau, Lionel Ott, and Fabio Ramos. Bayesian curiosity for efficient exploration in reinforcement learning. *arXiv preprint arXiv:1911.08701*, 2019.

Salomon Bochner et al. *Lectures on Fourier integrals*, volume 42. Princeton University Press, 1959.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*, 2018.

Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, 24, 2011.

Sayak Ray Chowdhury and Aditya Gopalan. Online learning in kernelized markov decision processes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3197–3205. PMLR, 2019.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Rodrigo de Lazcano, Kallinteris Andreas, Jun Jet Tai, Seungjae Ryan Lee, and Jordan Terry. Gymnasium robotics, 2024. URL http://github.com/Farama-Foundation/Gymnasium-Robotics.

Omar Darwiche Domingues, Pierre Ménard, Matteo Pirotta, Emilie Kaufmann, and Michal Valko. A kernel-based approach to non-stationary reinforcement learning in metric spaces. In *International Conference on Artificial Intelligence and Statistics*, pages 3538–3546. PMLR, 2021.

C. Daniel Freeman, Erik Frey, Anton Raichuk, Sertan Girgin, Igor Mordatch, and Olivier Bachem. Brax - a differentiable physics engine for large scale rigid body simulation, 2021. URL http://github.com/google/brax.

Karl Friston, Francesco Rigoli, Dimitri Ognibene, Christoph Mathys, Thomas Fitzgerald, and Giovanni Pezzulo. Active inference and epistemic value. *Cognitive Neuroscience*, 6 (4):187–214, 2015. doi: 10.1080/17588928.2015.1020053.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Junya Honda and Akimichi Takemura. An asymptotically optimal bandit algorithm for bounded support models. In *COLT*, pages 67–79. Citeseer, 2010.

Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. Vime: Variational information maximizing exploration. *Advances in neural information processing systems*, 29, 2016.

Carl Hvarfner, Erik Orm Hellsten, and Luigi Nardi. Vanilla bayesian optimization performs great in high dimensions. *arXiv preprint arXiv:2402.02229*, 2024.

Tommi S Jaakkola and David Haussler. Probabilistic kernel regression models. In *Seventh International Workshop on Artificial Intelligence and Statistics*. PMLR, 1999.

J Zico Kolter and Andrew Y Ng. Near-bayesian exploration in polynomial time. In *Proceedings of the 26th annual international conference on machine learning*, pages 513–520, 2009.

Robert Tjarko Lange. gymnax: A JAX-based reinforcement learning environment library, 2022. URL http://github.com/RobertTLange/gymnax.

Neil Lawrence, Matthias Seeger, and Ralf Herbrich. Fast sparse gaussian process methods: The informative vector machine. *Advances in neural information processing systems*, 15, 2002.

Haozhe Ma, Zhengding Luo, Thanh Vinh Vo, Kuankuan Sima, and Tze-Yun Leong. Highly efficient self-adaptive reward shaping for reinforcement learning. *arXiv preprint arXiv:2408.03029*, 2024.

Philippe Morere and Fabio Ramos. Bayesian RL for goal-only rewards. In *Conference on Robot Learning*, 2018.

Nikolay Nikolov, Johannes Kirschner, Felix Berkenkamp, and Andreas Krause. Information-directed exploration for deep reinforcement learning. *arXiv preprint arXiv:1812.07544*, 2018.

Ian Osband, Daniel Russo, and Benjamin Van Roy. (more) efficient reinforcement learning via posterior sampling. *Advances in Neural Information Processing Systems*, 26, 2013.

Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped dqn. *Advances in neural information processing systems*, 29, 2016.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pages 2778–2787. PMLR, 2017.

Fabien Pesquerel and Odalric-Ambrym Maillard. Imed-rl: Regret optimal learning of ergodic markov decision processes. *Advances in Neural Information Processing Systems*, 35:26363–26374, 2022.

Waris Radji. Pointax: Jax-native pointmaze environment, 2025. URL https://github.com/riiswa/pointax.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. *Advances in neural information processing systems*, 27, 2014.

Craig Saunders, Alexander Gammerman, and Volodya Vovk. Ridge regression learning algorithm in dual variables. 1998.

Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. A generalized representer theorem. In *International conference on computational learning theory*, pages 416–426. Springer, 2001.

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

Günther Schulz. Iterative berechung der reziproken matrix. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, 13 (1):57–59, 1933.

Burr Settles. Active learning literature survey. 2009.

Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

Alexander L Strehl and Michael L Littman. An analysis of model-based interval estimation for markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

Bhavya Sukhija, Stelian Coros, Andreas Krause, Pieter Abbeel, and Carmelo Sferrazza. Maxinforl: Boosting exploration in reinforcement learning through information gain maximization. *arXiv preprint arXiv:2412.12098*, 2024.

Danica J Sutherland and Jeff Schneider. On the error of random fourier features. *arXiv preprint arXiv:1506.02785*, 2015.

Richard S Sutton, Andrew G Barto, et al. Introduction to reinforcement learning, vol. 135, 1998.

Haoran Tang, Rein Houthooft, Davis Foote, Adam Stooke, OpenAI Xi Chen, Yan Duan, John Schulman, Filip DeTurck, and Pieter Abbeel. # exploration: A study of count-based exploration for deep reinforcement learning. *Advances in neural information processing systems*, 30, 2017.

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.

Michal Valko, Nathaniel Korda, Rémi Munos, Ilias Flaounas, and Nelo Cristianini. Finite-time analysis of kernelised contextual bandits. *arXiv preprint arXiv:1309.6869*, 2013.

Christopher Williams and Carl Rasmussen. Gaussian processes for regression. *Advances in neural information processing systems*, 8, 1995.

Max A Woodbury. *Inverting modified matrices*. Department of Statistics, Princeton University, 1950.

Zhitong Xu, Haitao Wang, Jeff M Phillips, and Shandian Zhe. Standard gaussian process is all you need for high-dimensional bayesian optimization. In *The Thirteenth International Conference on Learning Representations*, 2024.

Kai Yang, Jian Tao, Jiafei Lyu, and Xiu Li. Exploration and anti-exploration with distributional random network distillation. *arXiv preprint arXiv:2401.09750*, 2024.

Kenny Young and Tian Tian. Minatar: An atari-inspired testbed for thorough and reproducible reinforcement learning experiments. *arXiv preprint arXiv:1903.03176*, 2019.

Mingqi Yuan, Roger Creus Castanyer, Bo Li, Xin Jin, Wenjun Zeng, and Glen Berseth. Rlexplore: Accelerating research in intrinsically-motivated reinforcement learning. *arXiv preprint arXiv:2405.19548*, 2024.