STEERING: Stein Information Directed Exploration for Model-Based Reinforcement Learning

Souradip Chakraborty [1] **Amrit Singh Bedi** [1] **Alec Koppel** [2] **Mengdi Wang** [3] **Furong Huang** [1] **Dinesh Manocha** [1]

Abstract

Directed Exploration is a crucial challenge in reinforcement learning (RL), especially when rewards are sparse. Information-directed sampling (IDS), which optimizes the information ratio, seeks to do so by augmenting regret with information gain. However, estimating information gain is computationally intractable or relies on restrictive assumptions which prohibit its use in many practical instances. In this work, we posit an alternative exploration incentive in terms of the integral probability metric (IPM) between a current estimate of the transition model and the unknown optimal, which under suitable conditions, can be computed in closed form with the kernelized Stein discrepancy (KSD). Based on KSD, we develop a novel algorithm STEERING: **STE**in information dir**E**cted exploration for model-based **R**einforcement Learn**ING**. To enable its derivation, we develop fundamentally new variants of KSD for discrete conditional distributions. We further establish that STEERING archives sublinear Bayesian regret, improving upon prior learning rates of information-augmented MBRL. Experimentally, we show that the proposed algorithm is computationally affordable and outperforms several prior approaches.

1. Introduction

Exploring effectively is a major challenge in reinforcement learning (RL), particularly when the rewards are sparse (Rengarajan et al., 2022; Achiam & Sastry, 2017). Recent research using model-based reinforcement learning (MBRL) with intrinsic curiosity (Pathak et al., 2017a; Burda et al., 2018b; Pathak et al., 2019) has offered a potential solution, but its theoretical understanding is relatively immature. On the other hand, posterior sampling reinforcement learning (PSRL) offers an efficient framework for balancing exploration and exploitation that is conceptually well-substantiated (Osband et al., 2013; Osband & Van Roy, 2014). However, PSRL may struggle in scenarios where many trajectories are uninformative, due to, e.g., reward sparsity (Russo & Van Roy, 2014a).

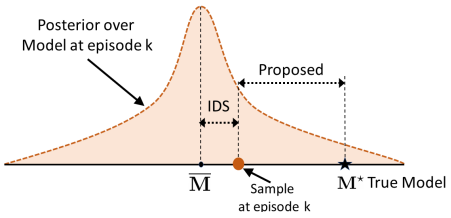


Figure 1. (Directed exploration) This figure illustrates that information-directed sampling (IDS) focuses on the distance between the current sample at episode k and the mean of the posterior. In contrast, we focus on the distance to the true model.

To incentivize exploration in MBRL, Lu & Van Roy (2019) proposed a design principle of information directed sampling (IDS), which optimizes the tradeoff between *regret* and *information*. Tight information-theoretic Bayesian regret bounds for IDS are derived in (Lu & Van Roy, 2019; Lu et al., 2021), under the specific choice of Dirichlet priors for the transition model, inspired by earlier work on bandits (Russo & Van Roy, 2014b). Recently, Hao & Lattimore (2022) alleviated any requirements on the prior based upon the development of a surrogate environment estimation procedure via rate-distortion theory.

Unfortunately, the existing IDS approaches face two main challenges (1) they are computationally intractable due to the need to estimate the information gain, (2) they do not induce exploration directed towards the optimal true transition dynamics. By *directed exploration*, we mean the algorithm moves in a direction toward the optimal transition dynamics rather than collecting all the information about the underlying environment, which is the focus of information gain-based exploration in IDS. The first challenge can be partially

[1]Department of Computer Science, University of Maryland, College Park, USA. [2]JP Morgan Chase AI Research, USA. [3]Department of Electrical Engineering, Princeton University/Deepmind, Princeton, NJ, USA. Correspondence to: Souradip Chakraborty <schakra3@umd.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

addressed by instead optimizing the evidence lower-bound (Achiam & Sastry, 2017), but ends up restricting focus to the posterior variance, which may be insufficiently informative about the underlying target distribution. This motivates us to pose the following question:

Can we develop a computationally tractable posterior sampling-based RL algorithm that exhibits efficient directed exploration with provable guarantees?

We provide an affirmative answer to this question by considering an alternative measure of information. That is, we propose *Stein information gain*, which is the integral probability metric (IPM) difference between the estimated and true (unknown) transition dynamics (Sriperumbudur et al., 2012), hence inducing *directed exploration*. Under the assumption that the transition model lies in the Stein class, we employ Stein's identity (Efron & Morris, 1973; James & Stein, 1992) to evaluate this IPM between the true (unknown) and estimated transitions in closed-form using kernelized Stein discrepancy (KSD) (Gorham & Mackey, 2015; Liu et al., 2016; Hawkins et al.). This is the key novelty that alleviates a major drawback of prior approaches that require evaluating mutual information. Thereby we introduce the Stein-information ratio, which may be seen as a modification of the information ratio in (Russo & Van Roy, 2018; Lu et al., 2021), and incentivizes exploration. We emphasize that our notion of KSD-based Stein information gain empowers us to evaluate the distance to the true transition dynamics. Doing so permits us to derive the best-known prior-free information-theoretic Bayesian regret bounds. Towards this end, we also develop the first KSD for conditional discrete distributions and employ it in tabular RL settings. Appendix A provides a detailed context of related works.

Contributions: Our main contributions are as follows.

- ▷ We formalize the setting of model-based episodic RL with Bayesian regret incorporating a notion of distance to the ground-truth MDP with KSD, and hence achieves directed exploration towards the optima.
- ▷ We introduce discrete conditional KSD (DSD) in tabular RL for the first time and use it to analyze distributional distance, which empowers us to evaluate the exploration incentive towards the true transition dynamics. Through this definition, we introduce a specific exploration-incentivized modification of posterior sampling, called `STEERING` (Algo. 1).
- ▷ We establish prior-free sublinear Bayesian regret bounds for `STEERING` and provide certain regularity conditions on the RKHS under which the regret can be further improved.
- ▷ We provide extensive experimental evidence for the proposed `STEERING` algorithm in sparse reward settings and show improvement via efficient directed exploration compared to all existing approaches.

2. Problem Formulation

We consider the problem of learning transition dynamics in an episodic finite-horizon time-homogeneous tabular Markov Decision Process (MDP) setting. We define the unknown MDP as a random variable $M := \{\mathcal{S}, \mathcal{A}, R, P, H, R_{\max}, \rho\}$, where $\mathcal{S}$ is finite state-space with $S = |\mathcal{S}|$, $\mathcal{A}$ is the finite action space with $A = |\mathcal{A}|$, and H is the episode length. Here, $P : \mathcal{S} \times \mathcal{A} \to \Delta_{\mathcal{S}}$ represents the transition dynamics for the state-action transitions and R is the rewards distribution where $\Delta_{\mathcal{S}}$ denotes the set of probability distributions over a finite set $\mathcal{S}$. After every episode of length H, state will reset according to the initial state distribution ρ. At time step $i \in [H]$ within an episode, the agent observes state $s_i \in \mathcal{S}$, selects action $a_i \in \mathcal{A}$ according to a stochastic policy $\pi : \mathcal{S} \to \Delta_{\mathcal{A}}$, receives reward $r_i \sim R(s_i, a_i)$ and transitions to a new state $s_{i+1} \sim P(\cdot|s_i, a_i)$. In this work, we consider M as a random process, as is often the case in Bayesian RL (Russo & Van Roy, 2014a).

Value Function and Bayesian Regret. For a given MDP M, the value for time step i is the reward accumulation during the episode, given by

$$V^M_{\pi,i}(s) = \mathbb{E}\left[\sum_{j=i}^{H} [\bar{r}^M(s_j, a_j)|s_i = s, a_j \sim \pi(\cdot|s_j)]\right], \quad (1)$$

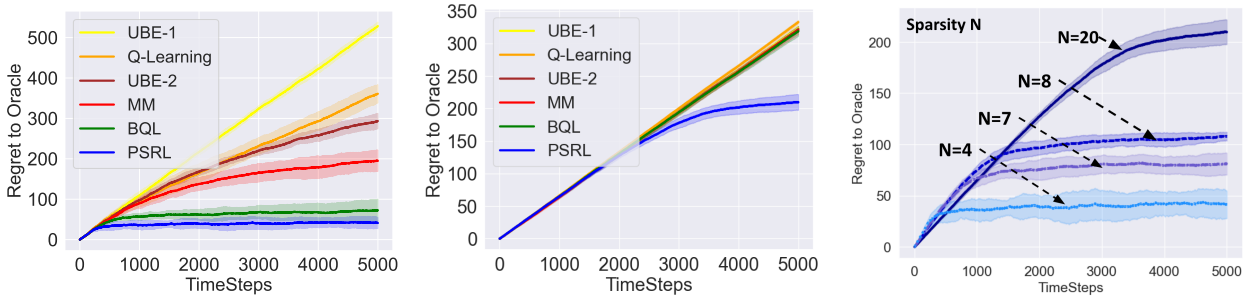
where j denotes the time-step within the episode and $\bar{r}^M(s, a) = \mathbb{E}_{r\sim R^M(s,a)}[r]$. Without loss of generality, we assume $|\bar{r}^M(s, a)| \leq R_{\max}$ for all $s \in \mathcal{S}, a \in \mathcal{A}$, which implies that $|V(s)| \leq HR_{\max}$, for all $s \in \mathcal{S}$. Next, for a given MDP M, an optimal policy π^* is

$$\pi^* = \operatorname*{argmax}_{\pi} V^M_{\pi,1}(s), \quad (2)$$

for all s and $i \in [H]$. We emphasize that since π^* is a function of M, it is also a random variable. An RL agent would interact with the environment over K number of episodes with policy $\{\pi^k\}_{k=1}^K$. The performance of the learning agent with respect to environment M can be quantified by *Bayesian regret*:

$$\mathfrak{BR}_K := \mathbb{E}\left[\sum_{k=1}^{K} (V^{M^*}_{1,\pi^*}(s^k_1) - V^{M^*}_{1,\pi^k}(s^k_1))\right], \quad (3)$$

where the expectation is with respect to the randomness in the policy π^k and the prior distribution of M^*, where M^* is the true MDP - a realization from the prior for an instantiation. Typically, one focuses on ensuring the sublinear growth of Bayesian regret in (3) as a way to quantify the learning performance of a given model-based estimate M_k



(a) Dense rewards. (b) Sparse rewards. (c) Variable sparsity (PSRL).

Figure 2. This figure compares the performance of various RL algorithms for Dense reward (Fig. 2(a)) and Sparse reward (Fig. 2(b)) DeepSea environment (DSE) (Osband & Van Roy, 2017a). As we move from dense to sparse reward settings, we note that the regret becomes almost linear for all algorithms except PSRL (Osband et al., 2013). We further show the performance degradation even for the PSRL algorithm with different sparsity levels in Fig. 2(c) achieved by varying N (denotes the number of states in DSE).

and the resultant policy $\pi^k := \text{argmax}_\pi V_\pi^{M_k}(s)$. Posterior sampling-based reinforcement learning (PSRL) operates this way (Osband & Van Roy, 2017a), as does upper-confidence bound (Ayoub et al., 2020), and Gittin's index (Edwards, 2019). However, regret [cf. (3)] alone may under-perform in contexts where the expected value function is a sparse function of the initial state, which can occur when the reward function is sparse, which is true for different applications (Weerakoon et al., 2022b; Chakraborty et al., 2022b).

To emphasize this point, we consider the DeepSea environment of Osband & Van Roy (2017a) and compare the performance of different existing methods in dense and sparse reward settings in Fig. 2. The performance degradation as we go from dense to sparse settings is evident (from Fig. 2(a) to Fig. 2(b)). We note that all the algorithms exhibit almost linear regret except PSRL, which establishes the efficient exploration aspect of PSRL. But as we investigate PSRL further for different sparsity levels in Fig. 2(c), we conclude that even PSRL suffers badly in very sparse reward settings. To deal with such challenges, information-directed sampling has been developed in the literature to jointly quantify value function sub-optimality with state space coverage.

Information Directed Sampling (IDS) augments PSRL by introducing the information ratio at episode k defined as

$$\Gamma_k(\pi, M^*) := \frac{\left(\mathbb{E}_k[V_{1,\pi^*}^{M^*}(s_1^k) - V_{1,\pi}^{M^*}(s_1^k)]\right)^2}{\mathbb{I}_k^\pi(M^*; \mathcal{H}_{k,H})}, \quad (4)$$

whose numerator is the Bayesian regret and the denominator is the information gain as used in (Russo & Van Roy, 2014a; Hao & Lattimore, 2022). The policy selection becomes $\pi_{\text{IDS}} = \arg\max_\pi \Gamma_k(\pi, M^*)$ which not only minimizes regret but does so while maximizing per unit information gain, which can yield efficient exploration. Following Hao & Lattimore (2022, Lemma A.1), to gain further insight, it is useful to rewrite the information gain as

$$\mathbb{I}_k^{\pi_{\text{TS}}^k}(M^*; \mathcal{H}_{k,H}) \quad (5)$$
$$= \sum_{h=1}^{H} \mathbb{E}_k \mathbb{E}_{\pi_{\text{TS}}^k}^{\overline{M}_k} \left[D_{\text{KL}} \left(P^{M^*}(\cdot|s_h^k, a_h^k) || P^{\overline{M}_k}(\cdot|s_h^k, a_h^k) \right) \right],$$

where $\mathbb{E}_{\pi_{\text{TS}}^k}^{\overline{M}_k}$ is taken with respect to s_h^k, a_h^k, and $\mathbb{E}_k$ is with respect to π and environment M. The expression in (5) makes it clear that incorporating $\mathbb{I}_k^\pi(M^*; \mathcal{H}_{k,H})$ into the objective motivates the agent to visit the state action region where $D_{\text{KL}}(P^{M^*} || P^{\overline{M}_k})$ is higher which acts as an intrinsic reward to explore state action pairs with high uncertainty.

2.1. Limitations of IDS

The ratio objective in (4) exhibits practical limitations related to the fact that the KL divergence in (5) cannot be evaluated. We next detail why this is so and how a modification can alleviate this issue.

(L1) Computational Intractability: A major challenge in prior approaches including (Hao & Lattimore, 2022; Lu & Van Roy, 2019) lies in an accurate estimation of mutual information. One of the first prior-free IDS analyses by Hao & Lattimore (2022) relies on constructing a covering set for KL divergence with cover radius $\epsilon = 1/KH$. This implies that the information gain term grows unbounded as the number of episodes K increases. Moreover, for practical implementation, to make the KL divergence in (5) tractable, Hao & Lattimore (2022) substitutes this quantity by its lower bound via Pinsker's inequality: $\mathbb{E}_k[D_{\text{KL}}(P^{M^*}(\cdot|s_h^k, a_h^k) || P^{\overline{M}_k}(\cdot|s_h^k, a_h^k))] \geq \sum_{s'} \text{Var}(P^{M^*}(s'|s_h^k, a_h^k))$. However, Ozair et al. (2019) shows that any high-confidence lower bound requires exponential samples in the mutual information, which is a critical concern. Hence even the variance lower bound with

Pinsker's inequality runs into the computational limits of estimating mutual information.

(L2) Not Truly A Directed Exploration: In the majority of the prior research on information-directed RL (Hao & Lattimore, 2022; Lu & Van Roy, 2019), the mutual information or KL divergence serves as the information-theoretic regularization or intrinsic curiosity to induce directed exploration to deal with the hard exploration challenges as detailed in (Russo et al., 2017). However, as we note from (5), since the mutual information must be substituted by the variance of the current estimate of the posterior distribution over M for tractability purposes, it only encourages the agent to explore trajectories with high variance. Ideally, we would want our exploration to be directed towards the true MDP M^* (see Fig. 1) from which the data (s, a, s') samples are collected in practice. A directed exploration towards M^* is crucial to avoid random wandering in the environment. For instance, consider a setting where we start with the strong belief prior, then, since the variance is already low, IDS based approach will not add any benefit on top of PSRL-based approaches.

To address the above limitations, we propose a novel notion of *Stein information gain* to achieve directed exploration in the next section.

3. Proposed Approach and Algorithm

In this work, we develop a novel Bayesian regret analysis that incorporates a notion of distance to the true optimal MDP and provides a computationally tractable alternative to the notion of information ratio in Hao & Lattimore (2022). Before presenting the proposed approach, let us discuss the technical development as follows.

3.1. Kernelized Stein Discrepancy

Integral probability metrics (IPM) have gained traction in Bayesian inference and generative modeling (Arjovsky et al., 2017) for their ability to quantify the merit of a given posterior distribution with respect to an unknown target without specifically having knowledge of that target. In particular, when one suitably assumes the class of posteriors over which the search is conducted to the *Stein class* (Liu et al., 2016), IPMs admit a closed-form evaluation in terms of Stein discrepancies. Please refer to Appendix B for detailed discussion and derivation. To this end, Liu et al. (2016) define kernel Stein discrepancy (KSD) between two distributions p and q as

$$\text{KSD}(p, q) = \mathbb{E}_{x,x' \sim p}\big[u_q(x, x')\big], \tag{6}$$

where $u_q(x, x')$ is the Stein kernel defined as

$$u_q(x, x') := s_q(x)^\top \kappa(x, x') s_q(x') + s_q(x)^\top \nabla_{x'} \kappa(x, x') + \nabla_x \kappa(x, x')^\top s_q(x') + trace(\nabla_{x,x'} \kappa(x, x')), \tag{7}$$

where $\kappa(x, x')$ is the base kernel (any positive definite kernel, for instance, Hamming Kernel for discrete rvs) The Stein kernel in (7) measures the similarity between two samples x and x', which comes from p, using the score function of q. For the setting in this work, we have $p = P^{M^*}$ (transition dynamics corresponding to true model M^*) and $q = P^{M_k}$ (transition dynamics corresponding to posterior $M_k \sim \phi(\cdot|\mathcal{H}_k)$). Interestingly, KSD empowers us to evaluate the distance $\text{KSD}(P^{M_k}, P^{M^*})$. In the next subsection, we present the main idea of this work.

3.2. Stein Information Directed Sampling

Now, after the introduction of KSD, we note that the distributional distance to the unknown target can be computed in closed form. We use this fact to address the limitations of IDS discussed in Sec. 2.1, and propose to replace information gain $\mathbb{I}_k^\pi(M^*; \mathcal{H}_{k,H})$ in the denominator of (4) with what we call *Stein information gain* $\mathbb{K}_k^\pi(M^*; \mathcal{H}_{k,H})$ given by

$$\mathbb{K}_k^\pi(M^*; \mathcal{H}_{k,H}) := \sum_{h=1}^{H} \mathbb{E}_k \mathbb{E}_\pi^{M^*} \left[\text{KSD}\left(P^{M_k}(\cdot|s_h^k, a_h^k), P^{M^*}(\cdot|s_h^k, a_h^k)\right)\right]. \tag{8}$$

where $\mathbb{E}_\pi^{M^*}$ is taken with respect to (s_h^k, a_h^k), and $\mathbb{E}_k$ is with respect to π and environment M_k. There are two main differences here as compared to information gain defined in (5). First, we use KSD to characterize if two transition models are close or not. Second, we use the distributional distance to true MDP M^* in (8) in contrast to posterior variance in (5). Hence, the ratio objective in (4) would modify to Stein information ratio as

$$\Gamma_k^{\text{KSD}}(\pi) := \frac{(\mathbb{E}_k[V_{1,\pi^*}^{M^*}(s_1^k) - V_{1,\pi}^{M^*}(s_1^k)])^2}{\mathbb{K}_k^\pi(M^*; \mathcal{H}_{k,H})}, \tag{9}$$

and we select $\pi_{\texttt{SIDS}}^k = \text{argmin}_\pi \Gamma_k^{\text{KSD}}(\pi)$. We remark that (8)-(9) are the point of departure from the existing IDS-based methods (Hao & Lattimore, 2022). Stein information gain term of (8) is different from the reduction in entropy (as in information gain) but instead characterizes the distributional distance to the true MDP transition dynamics P^{M^*}. So it forces the algorithm to move the model estimate towards the optimal than focusing on the coverage of space induced by information gain term in (4) (addressing L2). Another interesting aspect of the modified ratio objective in (9) is that it is computationally tractable due to the use of KSD and we no longer need to utilize Pinsker's inequality to approximate uncertainty via posterior variance which is unavoidable in IDS based approaches for practical implementation (Hao & Lattimore, 2022) (addressing L1) .

Unfortunately, although KSD is well-established, the pre-existing machinery (3.1) does not directly apply to our setting in (8). The impediments are twofold: firstly, (6) holds

for the continuous smooth densities, but our setting is tabular MDP; secondly, (6) applies to estimating unconditional target distributions, but in an MDP context, we require conditional distributions. To adapt Stein discrepancy to our setting, we need to first address both of these issues next.

3.3. Discrete Conditional KSD

In this subsection, we introduce the kernelized conditional discrete stein operator and *Discrete conditional kernelized Stein Discrepancy (DSD)* for analyzing the distributional distance between P^{M_k} where $M_k \sim \phi(\cdot|\mathcal{H}_k)$ and P^{M^*}. This is a unique contribution of this work which may be of independent interest in mathematical statistics. In tabular RL setting, if we are in state (s, a), then we know $s' \sim P^{M^*}(\cdot|s, a)$, and up to k^{th} episode, we collect samples in dictionary $\mathcal{D}_k := \{((s_1, a_1), s_1'), ((s_2, a_2), s_2'), \cdots ((s_k, a_k), s_k')\}$. Now, the objective is to derive DSD between $P^{M_k}(\cdot|s, a)$ and ground truth $P^{M^*}(\cdot|s, a)$ leveraging the recent literature on conditional independence testing with Kernel Stein's method (Jitkrittum et al., 2020) (defined only for continuous smooth densities). For simplicity and analysis in this subsection, let us denote the state-action pair $(s, a) \to x \in \mathcal{X} := \mathcal{S} \times \mathcal{A}$ and the corresponding next state $s' \to y \in \mathcal{S}$.

To write DSD, we start by defining the Stein operator (cf. Appendix B for unconditional case) as

$$\kappa_{M_k}((x, y), \cdot) = G(x, \cdot)\big[s_{P^{M_k}}(y)l(y, \cdot) - \triangle^* l(y, \cdot)\big], \tag{10}$$

where M_k signifies the dependence on P^{M_k}, function $l : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ is a positive definite kernels, and $G(x, \cdot)$ is a real-valued kernel associated with the RKHS $\mathcal{F}_k$ and explicitly defined in Proposition 3.1. Further in (10), we define score function $s_{P^{M_k}}(y)$ for P^{M_k} and $\triangle^*$ as the difference operator w.r.t inverse permutation denoted by $\wedge$ for the set $\mathcal{S}$. We denote $\triangle$ as the cyclic permutation $\vee$ for the set $\mathcal{S}$. For example, with $\mathcal{S} = \{+1, -1\}$, $\vee s = -s$, $\forall s \in \mathcal{S}$. On the other hand, inverse permutation satisfies $\vee(\wedge(s)) = \wedge(\vee(s)) = s$ (see (Yang et al., 2018) more details). Therefore, we expand the terms in (10) as

$$s_{P^{M_k}}(y)_i = \frac{\triangle_{y_i} P^{M_k}(y|x)}{P^{M_k}(y|x)} = 1 - \frac{P^{M_k}(\vee_i y)}{P^{M_k}(y|x)} \tag{11}$$

$$\triangle^*_{y_i} l(y, \cdot) = l(y, \cdot) - l(\wedge_i y, \cdot), \tag{12}$$

for $i = 1, 2, \cdots, S$. Next, we provide the definition of DSD between two conditional pmfs in Proposition 3.1.

Proposition 3.1. Let $G(x, \cdot) \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a real-valued kernel associated with the RKHS $\mathcal{F}_k$ and $l \colon \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ be positive definite kernels. Assume $G_x(x, x') := k(x, x')I$ for a positive definite kernel $k \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Then, DSD between $P^{M_k}(s'|s, a)$ and $P^{M^*}(s'|s, a)$ is given by

$$\text{DSD}(P^{M_k}, P^*) = \mathbb{E}_{[(x,y),(x',y')]}[\kappa_{M_k}((x, y), (x', y'))], \tag{13}$$

where (x, y) and (x', y') are samples from the joint distribution P^{M^*}. For simplicity of notations, we will denote $\text{DSD}(P^{M_k}, P^{M^*}) \to \text{DSD}(P^k)$ which quantifies the Stein distance of M_k from the true MDP M^*.

The proof of Proposition 3.1 is provided in Appendix C and we show that $\text{DSD}(P^{M^*}(\cdot|s, a)) = 0$ iff $P^{M_k}(\cdot|s, a) = P^{M^*}(\cdot|s, a)$. In the tabular RL setting, since both $x \in \mathcal{X}$ and $y \in \mathcal{S}$ are discrete random variables, we consider both $l(y, y') = \exp\{-H(y, y')\}$ and $k(x, x') = \exp\{-H(x, x')\}$ as the exponential Hamming kernel which is a positive definite kernel. However, the specific selection of kernels and kernel design in tabular RL is the scope of future work. We remark that since now DSD is defined, we will use DSD in place of KSD in all the places moving forward.

3.4. `STEERING`: Proposed Algorithm

Now, we are ready to present the proposed STEin information dirEcted sampling for model-based Reinforcement LearnING (`STEERING`) algorithm for tabular settings, summarized in Algorithm 1. We begin `STEERING` by assuming a prior over the transition, and the rewards model denoted by $\phi_{\mathcal{D}_1} = \{\mathcal{P}_{\mathcal{D}_1}, \mathcal{R}_{\mathcal{D}_1}\}$. At episode k, `STEERING` first samples a transition model P^{M_k} and rewards model R^{M_k} from the posterior distribution $\phi_{\mathcal{D}_k} = \{\mathcal{P}_{\mathcal{D}_k}, \mathcal{R}_{\mathcal{D}_k}\}$. It then optimizes the policy under these sampled models by minimizing the proposed Stein information ratio $\text{argmin}_\pi \Gamma_k^{\text{DSD}}(\pi)$ as in (9). This inner optimization procedure is a key aspect of `STEERING`, as it uses the Stein information to guide exploration. Finally, the agent interacts with the real environment using the resulting policy $\pi^k_{\texttt{IDS}}(a|s)$ to collect new samples at episode k and store them in $\mathcal{C}$. Instead of just appending $\mathcal{C}$ to dictionary $\mathcal{D}_k$, we propose to use an intelligent sample selection procedure (similar to SPMCMC (Chen et al., 2019)) on the collected samples in each episode k before updating the dictionary $\mathcal{D}_k$. This procedure selects new samples by minimizing their similarity to existing samples in the dictionary $\mathcal{D}_k$ through a local optimization procedure, as outlined in Appendix E starting from equation (38). This selection procedure helps to derive tighter convergence rates in the next section.

4. Regret Analysis

In this section, we derive the merits of incorporating Stein's method into the Bayesian regret of an MBRL method for the first time. We start by deriving our key result in Theorem 4.1, which connects Bayesian regret (cf. (3)) with Stein

Algorithm 1 STEERING: **STE**in information dir**E**cted sampling for model-based **R**einforcement Learn**ING**

Input : Episode length H, Total timesteps T, Dictionary $\mathcal{D}$, prior distribution $\phi = \{\mathcal{P}, \mathcal{R}\}$, policy $\pi(a|s)$, hyperparameter γ.
Initialization : Initialize dictionary $\mathcal{D}_1$ with random $\pi_0(a|s)$, posterior $\phi_{\mathcal{D}_1} = \{\mathcal{P}_{\mathcal{D}_1}, \mathcal{R}_{\mathcal{D}_1}\}$, policy $\pi_1(a|s)$.
for Episodes $k = 1$ to K **do**
  Sample a transition $P^{M_k} \sim \mathcal{P}_{\mathcal{D}_k}$ and reward model $R^{M_k} \sim \mathcal{R}_{\mathcal{D}_k}$ and initialize empty $\mathcal{C} = [\,]$
  Estimate the optimal policy by minimizing the *Stein information ratio* : $\pi^k_{\texttt{IDS}}(a|s) \leftarrow \operatorname{argmin}_\pi \Gamma^{\text{DSD}}_k(\pi, M^*)$ as in (9)
  Interact with the environment using the optimal policy $\pi^k_{\texttt{IDS}}(a|s)$ to gather and initialize empty $\mathcal{C} = \{s_{k,1}, a_{k,1}, r_{k,1}, \cdots, s_{k,H}, a_{k,H}, r_{k,H}\}$
  Select the subset of samples $\mathcal{C}' \in \mathcal{C}$ with least similarity to sample in samples in $\mathcal{D}_k$ in terms of Stein kernel
  Update dictionary $\mathcal{D}_{k+1} \leftarrow \mathcal{D}_k \cup \mathcal{C}'$ and update the posteriors $\mathcal{P}_{\mathcal{D}_k}, \mathcal{R}_{\mathcal{D}_k}$
end for

information gain (cf. (8)). Then, we analyze the evolution of the model-based estimates in terms of DSD [cf. Sec. 3.3], which decreases with the number of samples processed (Lemma 4.2). Next, we connect this bound to the Stein information ratio objective and the total Stein information gain in Lemma 4.3. Finally, we utilize Theorem 4.1, Lemma 4.2, and Lemma 4.3 to derive Bayesian regret in terms of state and action space cardinality in Theorem 4.4 . We also extend our analysis to regularized settings in Theorem 4.6. Next, we present our first Bayesian regret as follows.

Theorem 4.1. *(Stein Information Theoretic Regret) When Algorithm 1 is run for K episodes of horizon length H, it achieves the following Bayesian regret:*

$$\mathfrak{BR}_K \leq \sqrt{\mathbb{E}[\Gamma^*] K \sum_{k=1}^{K} \mathbb{E}[\mathbb{K}^\pi_k\left(M^*; \mathcal{H}_{k,H}\right)]}\,, \qquad (14)$$

where $\mathbb{K}^\pi_k\left(M^*; \mathcal{H}_{k,H}\right)$ is the Stein information gain (cf. (8)*) and Γ^* is the worst case Stein information ratio such that $\Gamma^{DSD}_k(\pi) \leq \Gamma^*$ for any $k \in K$ and π.*

The proof of Theorem 4.1 is provided in Appendix D. We call the regret in Theorem 4.1 as the *Stein information theoretic regret* because it upper bounds the Bayesian regret in terms of Stein information gain (cf. (8)), which is a DSD between the estimated model and the true model. This is the main point of departure as compared to information-theoretic regret derived in Osband & Van Roy (2017b); Hao & Lattimore (2022), which eventuates in substitution of the information gain by the posterior variance as its uncertainty quantifier due to the computational effort required to estimate information gain. By contrast, we consider this distributional distance instead in terms of integral probability metrics, which under some hypotheses, are computable as DSD (cf. Sec. 3.3). To the best of our knowledge, this is the first time this notion of distance to ground truth, which is typical of frequentist analysis of Bayesian methods, has been incorporated into the Bayesian regret.

Next, we proceed toward deriving an absolute upper bound on the regret in terms of S, A, and H. To achieve that, we present two intermediate results in Lemma 4.2-4.3.

Lemma 4.2. *(*DSD Convergence Rate) With Algorithm 1, we collect dictionary $\mathcal{D}_k$ for which it holds that

$$\mathbb{E}_k\left[\text{DSD}(P^k; \mathcal{D}_k)^2\right] = \mathcal{O}\left(\frac{S^2 A}{k}\right), \qquad (15)$$

for all k. Here $\text{DSD}(P^k) := \text{DSD}((P^{M^*}, P^{M^k})$ In (15), and $\text{DSD}(P^k; \mathcal{D}_k)^2$ denotes the empirical approximation using dictionary $\mathcal{D}_k$ of discretized conditional kernelized Stein discrepancy (cf. (13)) between P^{M_k} and true P^{M^*}. An improved bound of order $\mathcal{O}(\frac{SA}{k})$, can be derived under certain boundedness and regularity conditions on the RKHS.

The proof of Lemma 4.2 is provided in Appendix E. Lemma 4.2 establishes the convergence of the transition model estimation P^{M_k} to the true model P^{M^*}, which is an important result to prove next Lemma 4.3.

Lemma 4.3. For Algorithm 1, after K episodes of horizon length H, it holds that

(1) Stein information ratio (cf. (9)) is upper bounded as $\mathbb{E}[\Gamma^{\text{DSD}}_k(\pi)] \leq SAH^3$ for all k.

(2) Total Stein information gain (cf. (8)) is bounded as

$$\sum\nolimits_{k=1}^{K} \mathbb{E}[\mathbb{K}^\pi_k\left(M^*; \mathcal{H}_{k,H}\right)] \leq HS^2A(\log K). \quad \square$$

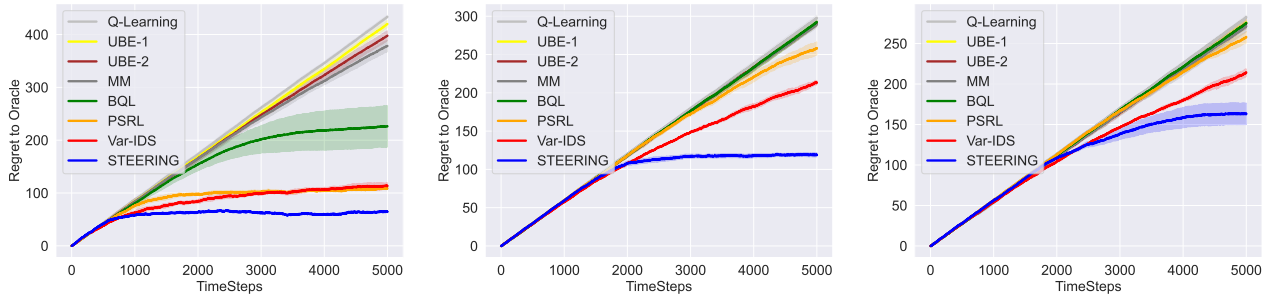
The proof is provided in Appendix F. The upper bounds established in Lemma 4.3 are crucial to specialize the general regret bounds developed in Theorem 4.1 in terms of state-action cardinalities, as follows in Theorem 4.4.

Theorem 4.4. *When Algorithm 1 is run for K episodes of horizon length H, it achieves the following performance in terms of Bayesian regret:*

$$\mathfrak{BR}_K = \tilde{\mathcal{O}}\left(\sqrt{H^4 S^3 A^2 K}\right), \qquad (16)$$

where $\tilde{\mathcal{O}}$ absorbs the log factors, S and A are the state and action space cardinalities, respectively.

The proof of Theorem 4.4 is provided in Appendix G. Theorem 4.4 states that STEERING achieves Bayesian regret,



(a) Low sparsity. (b) Medium sparsity. (c) High sparsity.

Figure 3. This figure compares the performance of STEERING on DeepSea enviornment (Osband & Van Roy, 2017a) against the existing RL baselines vanilla Q-learning with $\epsilon-$greedy action selection (Watkins & Dayan, 1992), Bayesian Q-learning (BQL) (Dearden et al., 1998), Uncertainty Bellman Equation (UBE) (O'Donoghue et al., 2017), Moment matching (MM) across Bellman equation (Markou & Rasmussen, 2019), Posterior Sampling RL (PSRL) (Osband et al., 2013), and IDS (Hao & Lattimore, 2022). We present results for three sparsity levels and observe STEERING outperforms existing baselines. Interestingly, the performance of STEERING is comparable (still much better shown in Fig. 3(a)-(b) to PSRL or IDS with low or medium sparsity, but for high sparsity in Fig. 3(c), STEERING significantly outperforms the other methods.

which is sublinear in terms of episode index K and the number of actions A, and linear in terms of the number of states S.

Remark 1 (Proof Sketch): Here, we provide insights regarding the regret proof of STEERING. One key factor in our analysis is the use of distributional directed sampling achieved via the integration of DSD and a point selection strategy inspired by SPMCMC (Chen et al., 2019). By utilizing an intelligent point selection method (cf. proof of Lemma 4.2 in Appendix E), we can establish convergence to underlying true distributions, as previously demonstrated in different contexts by (Koppel et al., 2021; Chen et al., 2019). Specifically, our point selection approach (Appendix E Eq. 38) $\inf_{(x,y)} \sum_{(x_i,y_i)\in\mathcal{D}_{k-1}} \kappa_{M_k}((x_i,y_i),(x,y))$ involves choosing a new sample (x,y) as the point that minimizes the similarity (in RKHS) to the current samples in the dictionary $\mathcal{D}_{k-1}$ resulting in directed exploration.

Remark 2 (Improved Regret): Here, we provided insights to obtain an improved version of regret provided in Theorem 4.4. We start by mentioning that the final bound in Theorem 4.4 depends upon the Stein information gain bound obtained in Lemma 4.3. We can obtain an improved bound on the Stein information gain term as summarized in Corollary 4.5.

Corollary 4.5. *Under additional assumptions on the structure of the Stein kernel, the Stein information gain upper bound (cf. Lemma 4.3) can be improved to* $HSA(\log K)$.

The details are provided in Appendix E and F. The result in Corollary 4.5 would lead to an improved regret bound of $\tilde{\mathcal{O}}(\sqrt{H^4S^2A^2K})$.

Regularized STEERING. We also consider a regularized Stein information gain-based sampling objective and optimize the policy as

$$\pi^k_{\texttt{r-IDS}} = \underset{\pi}{\operatorname{argmax}}\, \mathbb{E}_k[V^{M^*}_{1,\mu}(s^k_1)]) + \lambda \mathbb{K}^\pi_k(M^*; \mathcal{H}_{k,H}) \,, \tag{17}$$

where λ is the unknown regularization parameter. Next, we prove that the new regularized objective incurs the same Bayesian regret as the Stein information ratio in (9).

Theorem 4.6. *(Regularized Regret) When Algorithm 1 (after replacing step 5 with* (17)*) is run for K episodes of horizon length H, it achieves the following performance in terms of Bayesian regret* $\mathfrak{BR}_K \leq \sqrt{\frac{3}{2}\mathbb{E}[\Gamma^*]K\sum_{k=1}^K \mathbb{E}[\mathbb{K}^\pi_k(M^*;\mathcal{H}_{k,H})]}$, *for* $\lambda = \sqrt{K\mathbb{E}[\Gamma^*]/\sum_{k=1}^K \mathbb{E}[\mathbb{K}^\pi_k(M^*;\mathcal{H}_{k,H})]}$.

The proof of Theorem 4.6 is provided in Appendix H.

5. Experiments

In this section, we evaluate the performance of the proposed STEERING algorithm and compare it with other existing state-of-the-art algorithms. Since we are interested in developing algorithms for efficient directed exploration for an RL agent, we consider a challenging tabular sparse reward environment of DeepSea Exploration (DSE) introduced in Osband & Van Roy (2017a) (Fig. 4). The DSE environment tests the agent's capability of directed and sustained exploration. The agent starts from the left-most state and can swim left or right from each of the N states in the environment with near zero rewards everywhere except a reward of $r = 1$ only on a successful swim-right to $s = N$ (see Appendix I.1 for more details). Hence, increasing the number of states N induces more sparsity in the environment,

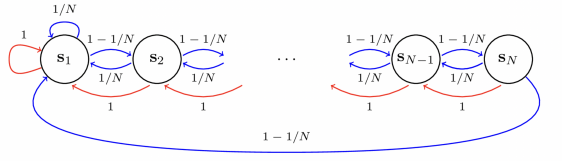


Figure 4. DeepSea Exploration Environment.

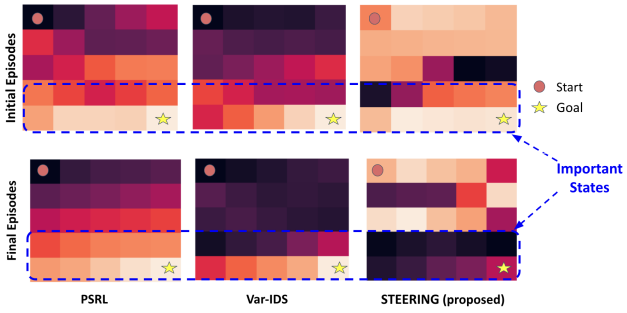
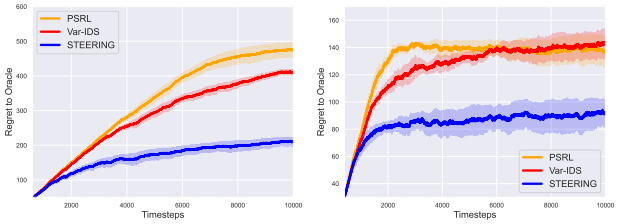


Figure 5. (Directed Exploration) We plot heatmaps showing the distribution of state-action visits in the initial and final episodes of training for PSRL, IDS, and STEERING. The results reveal that STEERING can effectively identify and visit the most important states indicated by the dark colors in the bottom row for STEERING. By important states, we mean the states closer to the goal state. In contrast, PSRL and IDS exhibit less directed exploration (see lighter colors in the bottom row for PSRL and IDS), with their heatmaps showing less concentration on important states. These findings demonstrate the superior performance of STEERING in guiding the exploration process towards optimality.

making it extremely hard for the agent to explore without directed exploration, as verified in Fig. 2(c). This motivates us to present improvements in the DSE environment. Next, we test STEERING on four different aspects of performance: (1) Regret to the Oracle, (2) Directed exploration, (3) Robustness to prior belief, and (4) Convergence to optimal value function. **(1) Regret to the Oracle:** First, in Fig. 3, we compare STEERING with other Bayesian/ non-Bayesian RL algorithms in terms of the cumulative regret accumulated with respect to an oracle agent following the optimal policy (refer to Appendix I for details). We present results in Fig. 3 for three different levels of sparsity: *low* ($N = 8$), *medium* ($N = 14$), and *high* ($N = 15$). The results validate our hypothesis that under highly sparse environments, general existing RL methods fail to explore efficiently, resulting in higher regret but STEERING outperforms.

(2) Directed Exploration by STEERING: To emphasize the nature of effective directed exploration provided by STEERING, we analyze its state-action space coverage in Fig. 5. We compare the heatmaps of state-action occupancy measure of the initial (top row) and final episodes (bottom row) for PSRL, IDS, and STEERING.



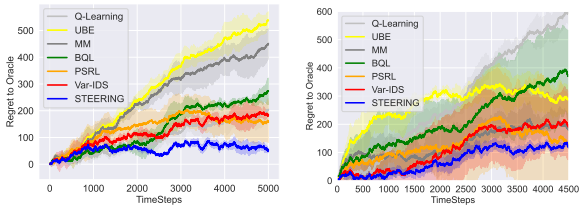
(a) Strong belief prior. (b) Weak belief prior.

Figure 6. We consider two belief levels: *weak* (higher variance) and *strong* (small variance) and compare the performances of PSRL, IDS, and STEERING. We note in Fig. 6(a), i.e., with strong prior belief, IDS provides a marginal benefit over PSRL, but STEERING is significantly better. This is due to the construction of STEERING based on the notion of distance to true MDP and not entirely relying on the posterior variance. Whereas IDS, on the other hand, for computational tractability, relies on posterior variance via Pinsker inequality. For weak prior belief in Fig. 6(b) also, STEERING performs favorably.

(3) Robustness to Prior Belief: The performance of Bayesian algorithms depends upon the prior, which might be a confident but mis-specified belief in practical settings. To test against such scenarios, we perform an ablation study in Fig. 6 to validate the robustness of STEERING to different levels of confidence of the prior belief over the MDP.

(4) Convergence to Optimal Value function: We perform additional experiments in Appendix J.1 to analyze and compare the convergence of the predicted $\hat{Q}$ values for by STEERING.

In Fig. 7, we also compare STEERING with baselines on WideNarrow MDP to validate its performance under factored posterior approximations and PriorMDP to validate its performance in general and practical environments without special structures (Markou & Rasmussen, 2019).



(a) WideNarrow MDP. (b) Prior MDP.

Figure 7. Performance comparison of STEERING and baseline algorithms in terms of Regret in two environments: WideNarrow MDP and PriorMDP.

6. Conclusions

Information-directed sampling (IDS) provides a way to induce exploration incentives into model-based reinforcement learning (MBRL). But IDS approaches suffer from computational tractability issues. To make such ratio-based approaches computationally tractable and efficient, we propose a novel measure to quantify directed exploration through a distributional distance to the optimal model via kernelized Stein discrepancy. To this end, we introduced a novel notion of Stein information gain and Stein information-directed sampling in MBRL. We theoretically established prior-free sublinear Bayesian regret bounds and experimentally demonstrate favorable performance in practice.

7. Acknowledgments

Chakraborty and Huang are supported by National Science Foundation NSF-IIS-FAI program, DOD-ONR-Office of Naval Research, DOD Air Force Office of Scientific Research, DOD-DARPA-Defense Advanced Research Projects Agency Guaranteeing AI Robustness against Deception (GARD), Adobe, Capital One, and JP Morgan faculty fellowships. Bedi and Manocha would like to acknowledge the support from Army Cooperative Agreement W911NF2120076 and Amazon Research Award 2022. Mengdi Wang acknowledges the support by NSF grants DMS-1953686, IIS-2107304, CMMI-1653435, ONR grant 1006977, and C3.AI.

References

Achiam, J. and Sastry, S. Surprise-based intrinsic motivation for deep reinforcement learning, 2017. URL https://arxiv.org/abs/1703.01732. 1, 2, 14

Ahmed, Z., Roux, N. L., Norouzi, M., and Schuurmans, D. Understanding the impact of entropy on policy optimization, 2018. URL https://arxiv.org/abs/1811.11214. 14

Amortila, P., Precup, D., Panangaden, P., and Bellemare, M. G. A distributional analysis of sampling-based reinforcement learning algorithms. In *International Conference on Artificial Intelligence and Statistics*, pp. 4357–4366. PMLR, 2020. 13

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017. 4

Ayoub, A., Jia, Z., Szepesvari, C., Wang, M., and Yang, L. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pp. 463–474. PMLR, 2020. 3

Bedi, A. S., Chakraborty, S., Parayil, A., Sadler, B. M., Tokekar, P., and Koppel, A. On the hidden biases of policy mirror ascent in continuous action spaces. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1716–1731. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/bedi22a.html. 14

Berlinet, A. and Thomas-Agnan, C. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011. 14

Borkar, V. S. and Meyn, S. P. Risk-sensitive optimal control for markov decision processes with monotone cost. *Mathematics of Operations Research*, 27(1):192–209, 2002. 13

Burda, Y., Edwards, H., Pathak, D., Storkey, A., Darrell, T., and Efros, A. A. Large-scale study of curiosity-driven learning, 2018a. URL https://arxiv.org/abs/1808.04355. 14

Burda, Y., Edwards, H., Pathak, D., Storkey, A., Darrell, T., and Efros, A. A. Large-scale study of curiosity-driven learning, 2018b. URL https://arxiv.org/abs/1808.04355. 1

Chakraborty, S., Bedi, A. S., Koppel, A., Sadler, B. M., Huang, F., Tokekar, P., and Manocha, D. Posterior coreset construction with kernelized stein discrepancy for model-based reinforcement learning. *arXiv preprint arXiv:2206.01162*, 2022a. 20

Chakraborty, S., Bedi, A. S., Koppel, A., Tokekar, P., and Manocha, D. Dealing with sparse rewards in continuous control robotics via heavy-tailed policies. *arXiv preprint arXiv:2206.05652*, 2022b. 3

Chakraborty, S., Bedi, A. S., Koppel, A., Tokekar, P., and Manocha, D. Dealing with sparse rewards in continuous control robotics via heavy-tailed policies, 2022c. 14

Chakraborty, S., Bedi, A. S., Koppel, A., Sadler, B. M., Huang, F., Tokekar, P., and Manocha, D. Posterior coreset construction with kernelized stein discrepancy for model-based reinforcement learning, 2023a. 13

Chakraborty, S., Weerakoon, K., Poddar, P., Tokekar, P., Bedi, A. S., and Manocha, D. Re-move: An adaptive policy design approach for dynamic environments via language-based feedback, 2023b. 14

Chen, W. Y., Barp, A., Briol, F.-X., Gorham, J., Girolami, M., Mackey, L., and Oates, C. Stein point markov chain monte carlo. In *International Conference on Machine Learning*, pp. 1011–1021. PMLR, 2019. 5, 7, 17, 20

Chowdhury, S. R. and Gopalan, A. Online learning in kernelized markov decision processes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3197–3205, 2019. 13

Chua, K., Calandra, R., McAllister, R., and Levine, S. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In *Advances in Neural Information Processing Systems*, pp. 4754–4765, 2018. 13

Dearden, R., Friedman, N., and Russell, S. Bayesian q-learning. In *Proceedings of the Fifteenth National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence*, AAAI '98/IAAI '98, pp. 761–768, USA, 1998. American Association for Artificial Intelligence. ISBN 0262510987. 7, 22, 30

Deisenroth, M. and Rasmussen, C. E. Pilco: A model-based and data-efficient approach to policy search. In *Proceedings of the 28th International Conference on machine learning (ICML-11)*, pp. 465–472, 2011. 13

Edwards, J. Practical calculation of gittins indices for multi-armed bandits, 2019. URL https://arxiv.org/abs/1909.05075. 3

Efron, B. and Morris, C. Stein's estimation rule and its competitors—an empirical bayes approach. *Journal of the American Statistical Association*, 68(341):117–130, 1973. 2

Eysenbach, B. and Levine, S. Maximum entropy rl (provably) solves some robust rl problems, 2021. URL https://arxiv.org/abs/2103.06257. 14

Fan, Y. and Ming, Y. Model-based reinforcement learning for continuous control with posterior sampling. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3078–3087. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/fan21b.html. 13

Gelfand, S. B. and Mitter, S. K. Recursive stochastic algorithms for global optimization in r^d. *SIAM Journal on Control and Optimization*, 29(5):999–1018, 1991. 14

Gorham, J. and Mackey, L. Measuring sample quality with stein's method. *Advances in Neural Information Processing Systems*, 28, 2015. 2, 14, 20

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290*, 2018. 13

Hao, B. and Lattimore, T. Regret bounds for information-directed reinforcement learning. *arXiv preprint arXiv:2206.04640*, 2022. 1, 3, 4, 6, 7, 14, 30

Hawkins, C., Koppel, A., and Zhang, Z. Online, informative mcmc thinning with kernelized stein discrepancy. *arXiv preprint arXiv:2201.07130*. 2, 17

Hawkins, C., Koppel, A., and Zhang, Z. Online, informative mcmc thinning with kernelized stein discrepancy, 2022. URL https://arxiv.org/abs/2201.07130. 20

James, W. and Stein, C. Estimation with quadratic loss. In *Breakthroughs in statistics*, pp. 443–460. Springer, 1992. 2

Janner, M., Fu, J., Zhang, M., and Levine, S. When to trust your model: Model-based policy optimization. In *Advances in Neural Information Processing Systems*, pp. 12498–12509, 2019. 13

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pp. 4863–4873, 2018. 13, 14

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pp. 2137–2143, 2020. 13, 14

Jitkrittum, W., Kanagawa, H., and Schölkopf, B. Testing goodness of fit of conditional density models with kernels, 2020. URL https://arxiv.org/abs/2002.10271. 5

Koppel, A., Pradhan, H., and Rajawat, K. Consistent online gaussian process regression without the sample complexity bottleneck. *Statistics and Computing*, 31(6):1–18, 2021. 7

Lai, T. L., Robbins, H., et al. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6 (1):4–22, 1985. 14

Lattimore, T. and Szepesvari, C. An information-theoretic approach to minimax regret in partial monitoring, 2019. URL https://arxiv.org/abs/1902.00470. 14

Li, L., Littman, M. L., and Walsh, T. J. Knows what it knows: A framework for self-aware learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pp. 568–575, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605582054. doi: 10.1145/1390156.1390228. URL https://doi.org/10.1145/1390156.1390228. 13

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning, 2015. URL https://arxiv.org/abs/1509.02971. 13

Littlestone, N. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. In *28th Annual Symposium on Foundations of Computer Science (sfcs 1987)*, pp. 68–77, 1987. doi: 10.1109/SFCS.1987.37. 13

Liu, J., Gu, X., and Liu, S. Policy optimization reinforcement learning with entropy regularization, 2019. URL https://arxiv.org/abs/1912.01557. 14

Liu, Q., Lee, J., and Jordan, M. A kernelized stein discrepancy for goodness-of-fit tests. In *International conference on machine learning*, pp. 276–284. PMLR, 2016. 2, 4, 14, 15, 20

Lu, X. and Van Roy, B. Information-theoretic confidence bounds for reinforcement learning, 2019. URL https://arxiv.org/abs/1911.09724. 1, 3, 4, 14

Lu, X., Van Roy, B., Dwaracherla, V., Ibrahimi, M., Osband, I., and Wen, Z. Reinforcement learning, bit by bit, 2021. URL https://arxiv.org/abs/2103.04047. 1, 2, 14

Markou, E. and Rasmussen, C. E. Bayesian methods for efficient reinforcement learning in tabular problems, 2019. URL https://github.com/stratisMarkou/sample-efficient-bayesian-rl. 7, 8, 21, 22, 23, 30

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. Playing atari with deep reinforcement learning, 2013. URL https://arxiv.org/abs/1312.5602. 13

O'Donoghue, B., Osband, I., Munos, R., and Mnih, V. The uncertainty bellman equation and exploration, 2017. URL https://arxiv.org/abs/1709.05380. 7, 22, 23, 30

Osband, I. and Van Roy, B. Model-based reinforcement learning and the eluder dimension. In *Advances in Neural Information Processing Systems*, pp. 1466–1474, 2014. 1, 13, 18, 20, 23

Osband, I. and Van Roy, B. Why is posterior sampling better than optimism for reinforcement learning? In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, pp. 2701–2710, International Convention Centre, Sydney, Australia, 2017a. PMLR. 3, 7, 13, 18, 20, 21, 22, 30

Osband, I. and Van Roy, B. Why is posterior sampling better than optimism for reinforcement learning? In *International conference on machine learning*, pp. 2701–2710. PMLR, 2017b. 6, 18, 23

Osband, I., Benjamin, V. R., and Daniel, R. (More) efficient reinforcement learning via posterior sampling. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pp. 3003–3011, USA, 2013. Curran Associates Inc. 1, 3, 7, 13, 18, 19, 20, 22, 23, 30

Osband, I., Van Roy, B., Russo, D. J., and Wen, Z. Deep exploration via randomized value functions. *Journal of Machine Learning Research*, 20(124):1–62, 2019. 14, 20

Ozair, S., Lynch, C., Bengio, Y., Oord, A. v. d., Levine, S., and Sermanet, P. Wasserstein dependency measure for representation learning, 2019. URL https://arxiv.org/abs/1903.11780. 3, 14

Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. Curiosity-driven exploration by self-supervised prediction, 2017a. URL https://arxiv.org/abs/1705.05363. 1

Pathak, D., Agrawal, P., Efros, A. A., and Darrell, T. Curiosity-driven exploration by self-supervised prediction, 2017b. URL https://arxiv.org/abs/1705.05363. 14

Pathak, D., Gandhi, D., and Gupta, A. Self-supervised exploration via disagreement, 2019. URL https://arxiv.org/abs/1906.04161. 1

Raginsky, M., Rakhlin, A., and Telgarsky, M. Non-convex learning via stochastic gradient langevin dynamics: a nonasymptotic analysis. In *Conference on Learning Theory*, pp. 1674–1703. PMLR, 2017. 14

Rengarajan, D., Vaidya, G., Sarvesh, A., Kalathil, D., and Shakkottai, S. Reinforcement learning with sparse rewards using guidance from offline demonstration, 2022. URL https://arxiv.org/abs/2202.04628. 1

Russo, D. and Van Roy, B. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39 (4):1221–1243, 2014a. 1, 2, 3

Russo, D. and Van Roy, B. An information-theoretic analysis of thompson sampling, 2014b. URL https://arxiv.org/abs/1403.5341. 1, 14

Russo, D. and Van Roy, B. Learning to optimize via information-directed sampling. *Operations Research*, 66(1):230–252, 2018. 2, 14

Russo, D., Van Roy, B., Kazerouni, A., Osband, I., and Wen, Z. A tutorial on thompson sampling. 2017. doi: 10.48550/ARXIV.1707.02038. URL https://arxiv.org/abs/1707.02038. 4, 14

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 13

Shyam, P., Jaśkowski, W., and Gomez, F. Model-based active exploration, 2018. URL https://arxiv.org/abs/1810.12162. 14

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research*, 11:1517–1561, 2010. 14

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., and Lanckriet, G. R. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012. 2

Valiant, L. G. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, nov 1984. ISSN 0001-0782. doi: 10.1145/1968.1972. URL https://doi.org/10.1145/1968.1972. 13

Watkins, C. J. C. H. and Dayan, P. Q-learning. *Machine Learning*, 8(3):279–292, May 1992. ISSN 1573-0565. doi: 10.1007/BF00992698. URL https://doi.org/10.1007/BF00992698. 7, 22, 30

Weerakoon, K., Chakraborty, S., Karapetyan, N., Sathyamoorthy, A. J., Bedi, A. S., and Manocha, D. Htron:efficient outdoor navigation with sparse rewards via heavy tailed adaptive reinforce algorithm, 2022a. 14

Weerakoon, K., Chakraborty, S., Karapetyan, N., Sathyamoorthy, A. J., Bedi, A. S., and Manocha, D. Htron:efficient outdoor navigation with sparse rewards via heavy tailed adaptive reinforce algorithm, 2022b. URL https://arxiv.org/abs/2207.03694. 3, 14

Yang, J., Liu, Q., Rao, V., and Neville, J. Goodness-of-fit testing for discrete distributions via stein discrepancy. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5561–5570. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/yang18c.html. 5

Yang, L. and Wang, M. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pp. 10746–10756. PMLR, 2020. 14

Zanette, A., Brandfonbrener, D., Brunskill, E., Pirotta, M., and Lazaric, A. Frequentist regret bounds for randomized least-squares value iteration. In *International Conference on Artificial Intelligence and Statistics*, pp. 1954–1964. PMLR, 2020. 14

Zimmert, J. and Lattimore, T. Connections between mirror descent, thompson sampling and the information ratio, 2019. URL https://arxiv.org/abs/1905.11817. 14

Appendix

Table of Contents

A. Detailed Context of Related Works

Model based and Model Free RL. Most RL algorithms fall into two categories: *model-free* (Schulman et al., 2017; Mnih et al., 2013; Haarnoja et al., 2018; Lillicrap et al., 2015) and *model-based* (Fan & Ming, 2021; Deisenroth & Rasmussen, 2011; Chua et al., 2018; Janner et al., 2019). In model-free approaches, the agent learns direct policy mapping from states to action with approximate dynamic programming methods. In contrast, in model-based approaches, an agent learns the approximate model of the environment itself and trains a policy under the learned dynamics. Recently, probabilistic model-based RL algorithms have shown superior performance in practice relative to their model-free counterparts, despite the strong conceptual guarantees for model-free approaches (Janner et al., 2019). We focus on model-based RL in this work.

Performance Measure. To understand why this may be so, it's important to assess the convergence criteria of RL methods: Probably approximate correct (PAC) bounds (Valiant, 1984), Frequentist regret (Jin et al., 2018; 2020), Bayesian regret (Osband et al., 2013), Mistake (MB) Bound (Littlestone, 1987), KWIK (Knows What It Knows) (Li et al., 2008) & convergence in various distributional metrics (Amortila et al., 2020; Borkar & Meyn, 2002; Chowdhury & Gopalan, 2019; Chakraborty et al., 2023a) all exist. A crucial challenge lies in deciding the optimal selection criteria to evaluate the algorithm's computational and statistical efficiency. Rather than comment on the specific merit of a particular convergence criterion as a motivation for our restriction of focus to Bayesian regret, we note a few of its salient attributes: it imposes minimal requirements on access to a generative model underlying state transitions (Osband & Van Roy, 2017a; Osband et al., 2013; Osband & Van Roy, 2014), and respects the inherent uncertainty associated with the optimal policy, rather than

supposing that it can be effectively captured by confidence sets based on a few moment-based estimates of the transition dynamics, which can lead to undesirable behavior in the presence of sparse rewards (Jin et al., 2020; Yang & Wang, 2020; Zanette et al., 2020). In addition, it has been observed in several practical scenarios that the performance degrades drastically with sparse rewards (Chakraborty et al., 2023b; Weerakoon et al., 2022a; Bedi et al., 2022; Chakraborty et al., 2022c).

Information-theoretic Approaches. To encapsulate the inherent uncertainty present in the optimal policy, one may augment notions of regret to quantify distance to the optimal occupancy measure or other information-theoretic quantities (Russo & Van Roy, 2018; Hao & Lattimore, 2022). That this is advantageous may be seen by honing in on the sparse reward setting: consider the traditional definition of regret $\mathbb{E}[V^*(s) - V^k(s)]$ for any k^{th} episode in an environment with near-zero rewards. Suppose one policy incorporates exploration based on a distributional estimate of the environment, whereas the other only consider moments of the distribution of returns, such as UCB (Lai et al., 1985). In this case, the traditional notion of regret may not encourage exploration in a way that yields increased state-space coverage, as the value distribution for this case would be near-null. This issue is well-documented in the bandit setting (Russo et al., 2017). Hence, there is an intrinsic motivation to consider augmentations of regret that are well-calibrated to the inherent uncertainty of the optimal policy. Motivated by (Russo & Van Roy, 2018; Hao & Lattimore, 2022), we consider convergence criteria from information theory and Bayesian inference to define an appropriate notion of regret. To understand the exact manner in which these modifications are incorporated into model-based RL (MBRL), we contrast them with the model-free setting. In such settings, incorporating exploration bonuses is well-established (Jin et al., 2020; Eysenbach & Levine, 2021; Liu et al., 2019; Ahmed et al., 2018), either in the form of augmenting the reward, the value function (Jin et al., 2018; Osband et al., 2019), or the policy gradient (Gelfand & Mitter, 1991; Raginsky et al., 2017). However, such methods sample uniformly with respect to a value or policy rather than in pursuit of reducing the estimation error to the optimal transition dynamics, which can yield spurious behavior (Weerakoon et al., 2022b; Shyam et al., 2018). By contrast, information-theoretic regularisation in model-based RL is an active area of research. Empirical advancements based on intrinsic curiosity (Burda et al., 2018a; Pathak et al., 2017b), i.e., modifying the sampling probabilities driven by uncertainty estimates in the transition dynamics or forward model prediction error, can improve performance in practice but lack conceptual guarantees.

Information Directed Sampling. To substantiate these approaches conceptually, the resultant algorithms have recently been rewritten in a way that their performance can be quantified by information-theoretic or Bayesian regret (Lu & Van Roy, 2019; Lu et al., 2021), under a specific choice of Dirichlet priors for the transition model, inspired by earlier work on bandits (Russo & Van Roy, 2014b). Extensions that alleviate any requirements on the prior for MBRL also exist based upon the development of a surrogate environment estimation procedure via rate-distortion theory (Hao & Lattimore, 2022). Unfortunately, the resultant algorithm requires estimating the information gain, which is generally intractable. This issue can be partially addressed by instead optimizing the evidence lower-bound (ELBO) (Achiam & Sastry, 2017), but exhibits exponential dependence on the mutual information with respect to the optimal occupancy measure (Ozair et al., 2019). Related approaches replace mutual information by Bregman divergence; however, this necessitates inverting a Fisher information matrix per step which can be computationally costly (Lattimore & Szepesvari, 2019; Zimmert & Lattimore, 2019). Hence, previous efforts to incorporate information-theoretic bonuses in MBRL either impose restrictive assumptions on the prior or yield computationally heavy objectives whose algorithmic solutions exhibit scalability problems.

B. Preliminaries: Kernelized Stein Discrepancy

Consider the notion of Integral probability metric to measure the deviation between the estimated distribution q and the unknown target distribution p defined as

$$d_{\mathcal{F}}(q, p) = \sup_{f \in \mathcal{F}} |\mathbb{E}_q[f(X)] - \mathbb{E}_p[f(X)]|, \tag{18}$$

where the supremum is over a class of real-valued test functions $f \in \mathcal{F}$. By adjusting the function class $\mathcal{F}$, we can recover the well-known metrics such as Total variation distance, Wasserstein distance (Sriperumbudur et al., 2010), etc. However, the major challenge in evaluating the IPM in (18) is that it requires an integration under the true distribution p which is intractable. A seminal idea to alleviate this issue is called Stein's method, which restricts the class of distributions $\mathcal{F}$ to functions such that $\mathbb{E}_p[f(X)] = 0$. Building upon this idea, (Liu et al., 2016) develops a tractable way to evaluate the IPM by restricting distributions to the Stein class, associated with a reproducing kernel Hilbert space (RKHS) over Stein kernels (Berlinet & Thomas-Agnan, 2011). In this case, the IPM can be evaluated in terms of the Stein kernel as the kernelized stein discrepancy (Gorham & Mackey, 2015). Stein's method provides a generalised framework for studying distributional distances and relies on the fact that two smooth densities $p(x)$ and $q(x)$ are identical iff they satisfy the Stein's identity given

by

$$\max_{f\in\mathcal{F}} \left(\mathbb{E}_p[s_q(x)f(x) + \nabla_x f(x)]\right)^2 = 0, \tag{19}$$

where $s_q(x)$ denotes the score function of $q(x)$ given by $s_q(x) = \nabla_x \log q(x)$. As an example, Stein's identity in (19) holds for smooth functions f lying in the Stein class of p. A function f is in the Stein class of p if it's smooth and satisfies $\int_x \nabla_x(f(x)p(x))dx = 0$. Hence, for any function f in the Stein class of p, we can say $\mathbb{E}_p[\mathcal{A}_p f(x)] = 0$ where $\mathcal{A}_p$ is the Stein operator of p which is a linear operator.

From here the Stein discrepancy between p and q is defined as (Liu et al., 2016)

$$\text{KSD}^2(p,q) = \max_{f\in\mathcal{F}} \left(\mathbb{E}_p[s_q(x)f(x) + \nabla_x f(x)]\right)^2, \tag{20}$$

where $\mathcal{F}$ is a class of smooth functions satisfying Stein's identity (19). However, the above definition is computationally intractable as it requires solving a complex variational optimization. To this end, (Liu et al., 2016) define a computationally tractable modification as

$$\text{KSD}(p,q) = \mathbb{E}_{x,x'\sim p}\big[u_q(x,x')\big], \tag{21}$$

where $u_q(x,x')$ is the Stein kernel defined as

$$\begin{aligned} u_q(x,x') := {} & s_q(x)^\top \kappa(x,x') s_q(x') + s_q(x)^\top \nabla_{x'}\kappa(x,x') \\ & + \nabla_x \kappa(x,x')^\top s_q(x') + Tr(\nabla_{x,x'}\kappa(x,x')), \end{aligned}$$

where $\kappa(x,x')$ is the base kernel.

For the setting in this work, we have $p = P^*$ (transition dynamics corresponding to true model M^*) and $q = P^{M_k}$ (transition dynamics corresponding to posterior $M_k \sim \phi(\cdot|\mathcal{H}_k)$). Interestingly, KSD empowers us to evaluate the distance $\text{KSD}(P^{M_k}, P^*)$.

C. Proof of Proposition 3.1

Proof. Let us first define the compact notation such that $G_x := G(x,\cdot)$ and $\xi_{P^{M^k}_{y|x}}(y,\cdot) := s_{P^{M^k}}(y)l(y,\cdot) - \triangle^* l(y,\cdot)$. Here, we derive our proposed DSD as defined in (13). Further, for simplicity of analysis, we denote $P^*(s'|s,a) \to P^*_{(y|x)}(y|x)$, $P^*(s,a) \to P^*_x(x)$ and $P^*(s,a,s') \to P^*_{(x,y)}(x,y)$, state-action pair $(s,a) \to x \in \mathcal{X} := \mathcal{S}\times\mathcal{A}$ and the corresponding next state $s' \to y \in \mathcal{S}$. Also, for simplicity of notation we denote $P^{M^*} \to P^*$ and $P^{M^k} \to P^k$. To start the proof, let us consider $\text{DSD}(P^k, P^*)$ and write

$$\begin{aligned} \text{DSD}(P^k,P^*) &= \left\|\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim P^*_{x,y}} G_{\mathbf{x}}\xi_{P^k_{y|x}}(\mathbf{y},\cdot)\right\|^2 && (22)\\ &= \left\langle \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim P^*_{(x,y)}} G_{\mathbf{x}}\xi_{P^k_{y|x}}(\mathbf{y},\cdot), \mathbb{E}_{(\mathbf{x}',\mathbf{y}')\sim P^*_{(x,y)}} G_{\mathbf{x}'}\xi_{P^k_{y'|x'}}(\mathbf{y}',\cdot)\right\rangle && (23)\\ &= \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim P^*_{(x,y)}}\mathbb{E}_{(\mathbf{x}',\mathbf{y}')\sim P^*_{(x,y)}} \left\langle G_{\mathbf{x}}\xi_{P^k_{y|x}}(\mathbf{y},\cdot), G_{\mathbf{x}'}\xi_{P^k_{y'|x'}}(\mathbf{y}',\cdot)\right\rangle && (24)\\ &= \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim P^*_{(x,y)}}\mathbb{E}_{(\mathbf{x}',\mathbf{y}')\sim P^*_{(x,y)}} \left\langle G_{\mathbf{x}'}G_{\mathbf{x}}\xi_{P^k_{y|x}}(\mathbf{y},\cdot), \xi_{P^k_{y'|x'}}(\mathbf{y}',\cdot)\right\rangle && (25)\\ &= \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim P^*_{(x,y)}}\mathbb{E}_{(\mathbf{x}',\mathbf{y}')\sim P^*_{(x,y)}}[k(\mathbf{x},\mathbf{x}')\kappa_P((\mathbf{x},\mathbf{y}),(\mathbf{x}',\mathbf{y}'))] && (26)\\ &= \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim P^*_{(x,y)}}\mathbb{E}_{(\mathbf{x}',\mathbf{y}')\sim P^*_{(x,y)}}[\kappa_k((\mathbf{x},\mathbf{y}),(\mathbf{x}',\mathbf{y}'))]. && (27) \end{aligned}$$

Here we have applied the reproducing property of kernels with the linearity of expectations to derive the equations. where, $G_{\mathbf{x}'}G_{\mathbf{x}} = G(x,x') = k(x,x')I$. This proves the derivation for an equivalent Stein operator for our scenario. Now, we need to show a version of Stein's identity to complete the derivation of our proposed Kernelized Conditional Discrete Stein Discrepancy. We note that

$$\begin{aligned} \mathbb{E}_{x,y\sim P^*_{(x,y)}}\kappa_k((x,y),\cdot) &= \mathbb{E}_{x,y\sim P^*_{(x,y)}} G_x\xi_{P^k_{y|x}}(\mathbf{y},\cdot) && (28)\\ &= \mathbb{E}_{x\sim P^*_x} G_x \mathbb{E}_{y\sim P^*_{(y|x)}}\xi_{P^k_{y|x}}(\mathbf{y},\cdot). && (29) \end{aligned}$$

Now, we show that if $P^k(y|x) = P^*(y|x)$, $\mathbb{E}_{y\sim P^*_{y|x}} \xi_{P^k_{y|x}}(\mathbf{y}, \cdot) = 0$ which proves an equivalent notion of Stein's identity for our Discrete conditional case. Replacing $P^k = P^*$.

$$\mathbb{E}_{y\sim P^*_{(y|x)}} \xi_{P^k_{y|x}}(\mathbf{y}, \cdot) = \mathbb{E}_{y\sim P^*_{(y|x)}}[s_{P^*_{(y|x)}}(y) l(y, \cdot) - \triangle^* l(y, \cdot)] \tag{30}$$

$$= \sum_y [s_{P^*_{(y|x)}}(y) l(y, \cdot) P^*(y|x) - \triangle^* l(y, \cdot) P^*(y|x)] \tag{31}$$

$$= \sum_y [\triangle P^*(y|x) l(y, \cdot) - \triangle^* l(y, \cdot) P^*(y|x)]. \tag{32}$$

Here the first equality holds by denoting $\xi_{P^k_{y|x}} = \left[s_{P^k}(y) l(y, \cdot) - \triangle^* l(y, \cdot)\right]$ from equation (10). Then expanding upon the expectation and replacing the expression of the score function from equation (11), we get the final expression. For each i we can write the first and second part of (32) from the definition of difference operators in equation (11) as

$$\sum_y [\triangle_{y_i} P^*(y|x) l(y, \cdot)] = \sum_y [l(y, \cdot) P^*(y|x) - l(y, \cdot) P^*(\vee_i y|x)], \tag{33}$$

$$\sum_y [\triangle^* l(y, \cdot) P^*(y|x)] = \sum_y [l(y, \cdot) P^*(y|x) - l(\wedge_i y, \cdot) P^*(y|x)]. \tag{34}$$

The two equations are equal since $\vee$ and $\wedge$ are inverse cyclic permutations on $\mathcal{S}$ with $\wedge_i(\vee_i y) = \vee_i(\wedge_i y) = y$ and hence substituting equation (33) into equation (32) we get $\mathbb{E}_{y\sim P^*(s'|s,a)} \xi_{y|x}(\mathbf{y}, \cdot) = 0$. So, this proves an equivalent notion of Stein's identity which completes the proof. □

D. Proof of Theorem 4.1

Proof. Let us start with the definition of Bayesian regret defined in (3) as follows

$$\mathfrak{BR}_K = \sum_{k=1}^K \mathbb{E}\left[\mathbb{E}_k\left[V^{M^*}_{1,\pi^*}(s^k_1) - V^{M^*}_{1,\pi^k_{\text{TS}}}(s^k_1)\right]\right], \tag{35}$$

where the inner expectation is over the posterior distribution, and the outer expectation is over the stochastic policy and environment M^*. For brevity of notation, let us define $\mathcal{R}_k := \mathbb{E}_k\left[V^{M^*}_{1,\pi^*}(s^k_1) - V^{M^*}_{1,\pi^k_{\text{TS}}}(s^k_1)\right]$. From here onwards, we use $\pi^k_{\text{TS}} \to \pi^k$ to represent the Posterior sampling policy for notation simplicity. Next, we introduce the Stein information ratio via multiplying and dividing by $\mathbb{K}^\pi_k(M^*; \mathcal{H}_{k,H})$ as follows

$$\mathfrak{BR}_K = \sum_{k=1}^K \mathbb{E}\left[\sqrt{\frac{(\mathcal{R}_k)^2}{\mathbb{K}^\pi_k(M^*; \mathcal{H}_{k,H})} \mathbb{K}^\pi_k(M^*; \mathcal{H}_{k,H})}\right].$$

After applying Cauchy–Schwartz inequality, using the linearity of expectations, and considering the definition of $\Gamma^{\text{DSD}}_k(\pi)$ in (9), we can get

$$\mathfrak{BR}_K \le \sqrt{\mathbb{E}\left[\sum_{k=1}^K \Gamma^{\text{DSD}}_k(\pi^k)\right]} \sqrt{\mathbb{E}\left[\sum_{k=1}^K \mathbb{K}^\pi_k(M^*; \mathcal{H}_{k,H})\right]}$$

$$= \sqrt{\mathbb{E}\left[\sum_{k=1}^K \mathbb{K}^\pi_k(M^*; \mathcal{H}_{k,H})\right] \sum_{k=1}^K \mathbb{E}[\Gamma^{\text{DSD}}_k(\pi^k)]}. \tag{36}$$

From definition of $\Gamma^{\text{DSD}}_k(\pi^k)$, we have $\Gamma^{\text{DSD}}_k(\pi^k) \le \Gamma^*$ for any $k \in [K]$. Hence, from (36), we can write (14). □

E. Proof of Lemma 4.2

Proof. We run a local optimization procedure by dividing the total number of samples H in an episode into batches of size Z with $Z' := \frac{H}{Z}$ batches and select Stein optimal points per batch using an SPMCMC style update. We begin the analysis by representing the dictionary till the k^{th} episode as $\mathcal{D}_k$ and we expand upon the definition of DSD (13) as

$$|\mathcal{D}_k|^2\text{DSD}^2(P^k;\mathcal{D}_k) = \sum_{(x_i,y_i)\in\mathcal{D}_k}\sum_{(x_j,y_j)\in\mathcal{D}_k}\kappa_k((x_i,y_i),(x_j,y_j)) \tag{37}$$

$$=|\mathcal{D}_{k-1}|^2\text{DSD}^2(P^{k-1};\mathcal{D}_{k-1}) + \sum_{z=1}^{Z'}\left[\kappa_k((x_k^z,y_k^z),(x_k^z,y_k^z)) + 2\sum_{(x_i,y_i)\in\mathcal{D}_{k-1}}\kappa_k((x_i,y_i),(x_k^z,y_k^z))\right]. \tag{38}$$

In the above expression, equality in (37) comes from the empirical definition of DSD, where κ_{M_k} denoted as κ_k is the Stein kernel which depend upon the score function of P^{M_k} (cf. (10)). Next, for each z, we select the sample (x_k^z, y_k^z) from $\mathcal{Y}_z := \{(x_k^l, y_k^l)\}_{l=1}^{Z}$ using an SPMCMC style local optimization procedure as in (Chen et al., 2019, Appendix A.1)). Now, from the SPMCMC-based selection, we can write

$$\kappa_k((x_k^z,y_k^z),(x_k^z,y_k^z)) + 2\sum_{(x_i,y_i)\in\mathcal{D}_{k-1}}\kappa_k((x_i,y_i),(x_k^z,y_k^z))$$

$$= \inf_{(x_k^z,y_k^z)\in\mathcal{Y}_m}\kappa_k((x_k^z,y_k^z),(x_k^z,y_k^z)) + 2\sum_{(x_i,y_i)\in\mathcal{D}_{k-1}}\kappa_k((x_i,y_i),(x_k^z,y_k^z))$$

$$\leq B^2 + 2\inf_{(x_k^z,y_k^z)\in\mathcal{Y}_z}\sum_{(x_i,y_i)\in\mathcal{D}_{k-1}}\kappa_k((x_i,y_i),(x_k^z,y_k^z)). \tag{39}$$

The inequality in (39) holds because we restrict our attention to regions for which it holds that $\kappa_k((x,y),(x,y)) \leq B^2$ for all $(x,y) \in \mathcal{Y}_k^z$ for all k and z. Utilizing the upper bound of (39) into the right hand side of (38), we get

$$|\mathcal{D}_k|^2\text{DSD}(P^k;\mathcal{D}_k)^2 \leq |\mathcal{D}_{k-1}|^2\text{DSD}(P^{k-1};\mathcal{D}_{k-1})^2 + Z'B^2 + 2\sum_{z=1}^{Z'}\inf_{\mathbf{h}_k^z\in\mathcal{Y}_z}\sum_{(x_i,y_i)\in\mathcal{D}_{k-1}}\kappa_k((x_i,y_i),(x_k^z,y_k^z)). \tag{40}$$

From the application of Theorem 5 (Hawkins et al.) for our formulation with H new samples in the dictionary.

$$2\inf_{(x_k^m,y_k^m)\in\mathcal{Y}_m}\sum_{(x_i,y_i)\in\mathcal{D}_{k-1}}\kappa_k((x_i,y_i),(x_k^m,y_k^m)) \leq r_k\|f_k\|_{\mathcal{K}_0}^2 + \frac{\text{DSD}(P^{k-1};\mathcal{D}_{k-1})^2}{r_k}, \tag{41}$$

for any arbitrary constant $r_k > 0$ and any $f_k = \sum\kappa_k(x_i,y_i),\cdot)$ which lies in RKHS spanned by the kernel $\kappa_k((x_i,y_i),(x_j,y_j))$ can be trivially upper-bounded as $\|f_k\|_{\mathcal{K}_0}^2 \leq C_0S^2A$, for any choice of kernel where C_0 is a positive constant. Hence, we can use the upper bound in (41) to the right hand side of (40), to obtain

$$|\mathcal{D}_k|^2\text{DSD}(P^k;\mathcal{D}_k)^2 \leq |\mathcal{D}_{k-1}|^2\left(1+\frac{Z'}{r_k}\right)\text{DSD}(P^{k-1};\mathcal{D}_{k-1})^2 + Z'(B^2 + r_kC_0S^2A). \tag{42}$$

Next, we divide the both sides by $|\mathcal{D}_k|^2 = (|\mathcal{D}_{k-1}| + Z')^2$ to obtain

$$\text{DSD}(P^k;\mathcal{D}_k)^2 \leq \frac{|\mathcal{D}_{k-1}|^2}{(|\mathcal{D}_{k-1}|+Z')^2}\left(1+\frac{Z'}{r_k}\right)\text{DSD}(P^{k-1};\mathcal{D}_{k-1})^2 + \frac{Z'\left(B^2 + r_kC_0S^2A\right)}{(|\mathcal{D}_{k-1}|+Z')^2} \tag{43}$$

It is interesting that a novel aspect in analysis lies in establishing a recursive relationship for the DSD amongst iterations which eventually paves the way to establish the DSD convergence results. After unrolling the recursion in (43), we can write

$$\text{DSD}(P^k;\mathcal{D}_k)^2 \leq \sum_{i=1}^{k}\left(\frac{Z'\left(B^2 + r_iC_0S^2A\right)}{(|\mathcal{D}_{i-1}|+Z')^2} + \epsilon_i\right)\left(\prod_{j=i}^{k-1}\frac{|\mathcal{D}_j|}{|\mathcal{D}_j|+Z'}\right)^2\left(\prod_{j=i}^{k-1}\left(1+\frac{Z'}{r_{j+1}}\right)\right). \tag{44}$$

Applying the log-sum exponential bound $\prod_{j=i}^{k-1}\left(1+\frac{Z'}{r_{j+1}}\right) \leq \exp\left(Z'\sum_{j=1}^{n}\frac{1}{r_j}\right)$, we can write (44) as

$$\text{DSD}(P^k;\mathcal{D}_k)^2 \leq \exp\left(Z'\sum_{j=1}^{k}\frac{1}{r_j}\right)\sum_{i=1}^{k}\left(\frac{Z'(B^2+r_iC_0S^2A)}{\left(|\mathcal{D}_{i-1}|+Z'\right)^2}+\right)\left(\prod_{j=i}^{k-1}\frac{|\mathcal{D}_j|}{|\mathcal{D}_j|+Z'}\right)^2. \tag{45}$$

Next, we consider the inequality in (45). By replacing, $r_j = \frac{k}{Z'}$, we get rid of the constant exponential term and obtain

$$\begin{aligned}\text{DSD}(P^k;\mathcal{D}_k)^2 &\leq \sum_{i=1}^{k}\left(\frac{Z'(B^2+r_iC_0S^2A)}{\left(|\mathcal{D}_{i-1}|+Z'\right)^2}\right)\left(\prod_{j=i}^{k-1}\frac{|\mathcal{D}_j|}{|\mathcal{D}_j|+Z'}\right)^2 \\ &= \sum_{i=1}^{k}\left(\frac{Z'(B^2+r_iC_0S^2A)}{\left(|\mathcal{D}_{k-1}|+Z'\right)^2}\right)\left(\prod_{j=i}^{k-1}\frac{|\mathcal{D}_j|}{|\mathcal{D}_{j-1}|+Z'}\right)^2,\end{aligned} \tag{46}$$

where Equation (46) corresponds to the sampling error and represents the bias incurred at each step of the SPMCMC point selection scheme. The equality in the second line holds by rearranging the denominators in the multiplication and pulling $\left(|\mathcal{D}_{k-1}|+Z'\right)^2$ inside the first term. Next, from the fact that $|\mathcal{D}_j| = |\mathcal{D}_{j-1}| + Z'$ which implies that the product will be 1, we can upper bound the right hand side of (46) as follows

$$\text{DSD}(P^k;\mathcal{D}_k)^2 \leq \sum_{i=1}^{k}\left(\frac{Z'(B^2+r_iC_0S^2A)}{\left(|\mathcal{D}_{k-1}|+Z'\right)^2}\right). \tag{47}$$

From the dictionary update, we note that $|\mathcal{D}_{k-1}| + Z' = |\mathcal{D}_k| = \mathcal{O}(k)$, which implies that $1/\left(|\mathcal{D}_{k-1}|+Z'\right)^2 = 1/k^2$, which we utilize in the right hand side of (47) to write

$$\text{DSD}(P^k;\mathcal{D}_k)^2 \leq \sum_{i=1}^{k}\left(\frac{Z'(B^2+r_iC_0S^2A)}{k^2}\right). \tag{48}$$

We note that the above bound holds for any given M. And hence we can conclude that after taking an expectation over posterior $M \sim \phi(\cdot|\mathcal{D}_k)$, it holds that

$$\mathbb{E}_k\left[\text{DSD}(P^k;\mathcal{D}_k)^2\right] \leq \sum_{i=1}^{k}\left(\frac{Z'(B^2+r_iC_0S)}{k^2}\right) = \mathcal{O}\left(\frac{S^2A}{k}\right), \tag{49}$$

where we absorb the constants into $\mathcal{O}(\cdot)$ notation. Hence proved. We can improve the above bound to $\mathcal{O}\left(\frac{SA}{k}\right)$ for kernels that satisfy some factorization properties. For instance, the above result holds for the Kronecker delta kernel where $\kappa((x,y),(x',y')) = \delta((x,y),(x',y'))$ which implies $k((x,y),(x',y')) = \delta(x,x')\delta(y,y')$, and leads to improved upper-bound in DSD of order $\mathcal{O}(\frac{SA}{k})$. This result would also hold for the bounded function $\|G(x,\cdot)\| \leq C$ (cf. (10)) for some constant C. This would help us to upper-bound the term $\|f_k\|^2_{\mathcal{K}_0} \leq C_0SA$ and subsequently follow the same steps ((41)-(49)) to get the final bound of $\mathcal{O}\left(\frac{SA}{k}\right)$.

□

F. Proof of Lemma 4.3

Proof. **Proof of statement (1):** We start by analyzing the regret decomposition of the value function as follows

$$\mathbb{E}_k\left[V^{M^*}_{1,\pi^*}(s^k_1) - V^{M^*}_{1,\pi^k}(s^k_1)\right] = \underbrace{\mathbb{E}_k\left[V^{M^*}_{1,\pi^*}(s^k_1) - V^{M^k}_{1,\pi^k}(s^k_1)\right]}_{I_1} + \underbrace{\mathbb{E}_k\left[V^{M^k}_{1,\pi^k}(s^k_1) - V^{M^*}_{1,\pi^k}(s^k_1)\right]}_{I_2}. \tag{50}$$

From the probability matching principle of PSRL (Osband et al., 2013; Osband & Van Roy, 2014; 2017b;a) we have $\mathbb{E}_k\left[V^{M^*}_{1,\pi^*}(s^k_1) - V^{M^k}_{1,\pi^k}(s^k_1)\right] = 0$ conditioned on the history D_k. At the start of each episode M^*, M^k are identically

distributed as detailed in (Osband et al., 2013), hence $I_1 = 0$

Upper Bound on I_2

Now, we derive the upper bound on I_2

$$\Delta_h^k(s,a) := \mathbb{E}_{s'\sim P^{M^k}(\cdot|s,a)}[V_{h+1,\pi^k}^{M^k}(s')] - \mathbb{E}_{s'\sim P^{M^*}(\cdot|s,a)}[V_{h+1,\pi^k}^{M^k}(s')]\,. \tag{51}$$

Now, using the definition in (51), we can write I_2 as

$$I_2 = \mathbb{E}_k\left[\sum_{h=1}^{H}\mathbb{E}_{\pi^k}^{M^*}\left[\Delta_h^k(s_h^k,a_h^k)\right]\right] \tag{52}$$

$$= \sum_{h=1}^{H}\mathbb{E}_k\left[\sum_{(s,a)} d_{h,\pi^k}^{M^*}(s,a)\Delta_h^k(s,a)\right], \tag{53}$$

$$= \sum_{h=1}^{H}\mathbb{E}_k\left[\sum_{(s,a)} \frac{d_{h,\pi^k}^{M^*}(s,a)}{(\mathbb{E}_k[d_{h,\pi^k}^{M^*}(s,a)])^{1/2}}(\mathbb{E}_k[d_{h,\pi^k}^{M^*}(s,a)])^{1/2}\Delta_h^k(s,a)\right]. \tag{54}$$

The equality in (53) holds by introducing the state action occupancy measure, and in (54), we divide and multiply with $d_{h,\pi^k}^{M^*}(s,a) > 0$. From the Cauchy–Schwartz inequality and using the fact that $d_{h,\pi^k}^{M^*}(s,a) \leq 1$, we can write

$$\begin{aligned} I_1 &\leq \left(\sum_{h=1}^{H}\mathbb{E}_k\left[\sum_{(s,a)}\frac{(d_{h,\pi^k}^{M^*}(s,a))^2}{\mathbb{E}_k[d_{h,\pi^k}^{M^*}(s,a)]}\right]\right)^{1/2}\left(\sum_{h=1}^{H}\mathbb{E}_k\left[\sum_{(s,a)}\mathbb{E}_k[d_{h,\pi^k}^{M^*}(s,a)](\Delta_h^k(s,a))^2\right]\right)^{1/2} \\ &\leq \left(\sum_{h=1}^{H}\mathbb{E}_k\left[\sum_{(s,a)}\frac{d_{h,\pi^k}^{M^*}(s,a)}{\mathbb{E}_k[d_{h,\pi^k}^{M^*}(s,a)]}\right]\right)^{1/2}\left(\sum_{h=1}^{H}\mathbb{E}_k\left[\sum_{(s,a)}\mathbb{E}_k[d_{h,\pi^k}^{M^*}(s,a)](\Delta_h^k(s,a))^2\right]\right)^{1/2}. \end{aligned} \tag{55}$$

First, from the first part in the right-hand side of (55), we note that

$$\sum_{h=1}^{H}\mathbb{E}_k\left[\sum_{(s,a)}\frac{d_{h,\pi^k}^{M^*}(s,a)}{\mathbb{E}_k[d_{h,\pi^k}^{M^*}(s,a)]}\right] = \sum_{h=1}^{H}\left[\sum_{(s,a)}\frac{\mathbb{E}_k[d_{h,\pi^*}^{M^*}(s,a)]}{\mathbb{E}_k[d_{h,\pi^k}^{M^*}(s,a)]}\right] \leq HSA\,. \tag{56}$$

Then, from the second part on the right-hand side of (55), we note that given $\mathcal{D}_k$, we have $d_{h,\pi^k}^{M^*}(s,a)$ and $\Delta_h^k(s,a)$ independent. This implies

$$\begin{aligned} \mathbb{E}_k\left[\sum_{(s,a)} d_{h,\pi^k}^{M^*}(s,a)\right]\mathbb{E}_k\left[(\Delta_h^k(s,a))^2\right] &= \mathbb{E}_k\left[\sum_{(s,a)} d_{h,\pi^k}^{M^*}(s,a)(\Delta_h^k(s,a))^2\right] \\ &= \mathbb{E}_k\left[\mathbb{E}_{\pi^k}^{M^*}\left[(\Delta_h^k(s_h^k,a_h^k))^2\right]\right], \end{aligned} \tag{57}$$

From the definition in (51), we can write

$$\begin{aligned} &\mathbb{E}_k\left[\mathbb{E}_{\pi^k}^{M^*}\left[(\Delta_h^k(s_h^k,a_h^k))^2\right]\right] \\ &= H^2\mathbb{E}_k\left[\mathbb{E}_{\pi^k}^{M^*}\left[\left(\mathbb{E}_{s'\sim P^{M^k}(\cdot|s_h^k,a_h^k)}\left[V_{h+1,\pi^k}^{M^k}(s')/H\right] - \mathbb{E}_{s'\sim P^{M^*}(\cdot|s_h^k,a_h^k)}\left[V_{h+1,\pi^*}^{M^k}(s')/H\right]\right)^2\right]\right] \\ &= H^2\mathbb{E}_k\left[\mathbb{E}_{\pi^k}^{M^*}\|P^{M^k}(\cdot|s_h^k,a_h^k) - P^{M^*}(\cdot|s_h^k,a_h^k)\|^2\right]. \end{aligned} \tag{58}$$

Next, combining (56), (57), and (58), can upper bound I_2 in (55) as

$$I_2 \leq \sqrt{H^3SA\sum_{h=1}^{H}\mathbb{E}_k\left[\mathbb{E}_{\pi^k}^{M^*}\|P^{M^k}(\cdot|s_h^k,a_h^k) - P^{M^*}(\cdot|s_h^k,a_h^k)\|^2\right]}. \tag{59}$$

Next, we take square on both sides and introduce the Kernelized Stein discrepancy between probability measures P^M and P^{M^*} by upper-bounding the total variation norm in equation (59) [1] (Gorham & Mackey, 2015) which is a clear departure from the traditional Bayesian regret analysis as in (Osband et al., 2013; Osband & Van Roy, 2017a; 2014; Osband et al., 2019). to obtain

$$I_2^2 \leq H^3SA \sum_{h=1}^{H} \mathbb{E}_k \left[\mathbb{E}_{\pi^k}^{M^*} [DSD^2(P^{M^k}(\cdot|s_h^k, a_h^k), P^{M^*}(\cdot|s_h^k, a_h^k))] \right]$$
$$\leq H^3SA \sum_{h=1}^{H} \mathbb{E}_k \left[\mathbb{E}_{\pi^k}^{M^*} [DSD^2(P^k(\cdot|s_h^k, a_h^k))] \right]. \tag{60}$$

The second line in the equation is due to the definition of KSD which requires only the unnormalized density/pmf of one and samples from the other. Also, we add the second line for notational consistency Next, from the definition in (8), we can finally write

$$I_2^2 \leq H^3SA \cdot \mathbb{K}_k^{\pi} \left(M^*; \mathcal{H}_{k,H} \right). \tag{61}$$

From (50), we can write

$$\left(\mathbb{E}_k \left[V_{1,\pi^k}^{M^k}(s_1^k) - V_{1,\pi^k}^{M^*}(s_1^k) \right] \right)^2 \leq H^3SA \cdot \mathbb{K}_k^{\pi} \left(M^*; \mathcal{H}_{k,H} \right). \tag{62}$$

Since $I_1 = 0$, as detailed above, adding that to equation (62), we get

$$\left(\mathbb{E}_k \left[V_{1,\pi^*}^{M^*}(s_1^k) - V_{1,\pi^k}^{M^*}(s_1^k) \right] \right)^2 \leq H^3SA \cdot \mathbb{K}_k^{\pi} \left(M^*; \mathcal{H}_{k,H} \right). \tag{63}$$

Hence, from equation (63) and using the definition of Stein information ratio in (9) and inequality in (62), it is clear that we can upper bound $\mathbb{E}[\Gamma_k^{DSD}]$ as $\mathbb{E}[\Gamma_k^{DSD}] \leq SAH^3$. Now from the definition of $\Gamma^* \leq \Gamma_k^{DSD}(\pi)$ we have $\Gamma^* \leq SAH^3$ that which completes the proof.

Proof of statement (2): In this section, we derive an upper-bound on the total Stein information for the K episodes given by $\sum_{k=1}^{K} \mathbb{E}[\mathbb{K}_k^{\pi} \left(M; \mathcal{H}_{k,H} \right)]$ in order to compute the second term in the equation (36). We start by deriving an upper bound for $\mathbb{E}[\mathbb{K}_k^{\pi} \left(M; \mathcal{H}_{k,H} \right)]$ with an SPMCMC style local optimization method proposed originally in (Chen et al., 2019) and later used in sequential decision-making scenarios by (Chakraborty et al., 2022a; Hawkins et al., 2022).
We start with the definition in (8) to write

$$\mathbb{E}_k[\mathbb{K}_k^{\pi} \left(M^*; \mathcal{H}_{k,H} \right)] = \sum_{h=1}^{H} \mathbb{E} \left[DSD^2(P^k(\cdot|s_h^k, a_h^k) \right]. \tag{64}$$

From the upper bound in Lemma 4.3, we can write

$$\mathbb{E}[\mathbb{K}_k^{\pi} \left(M^*; \mathcal{H}_{k,H} \right)] \leq \sum_{h=1}^{H} \mathcal{O} \left(\frac{S^2A}{k} \right) = \mathcal{O} \left(\frac{HS^2A}{k} \right). \tag{65}$$

Finally, we take summation over k and obtain the upper bound as

$$\sum\nolimits_{k=1}^{K} \mathbb{E}[\mathbb{K}_k^{\pi} \left(M^*; \mathcal{H}_{k,H} \right)] \leq HS^2A \int_1^K \frac{1}{x} dx = HS^2A(\log K). \tag{66}$$

Hence Proved. With the Corollary 4.5, the following upper-bound can be improved by an order of S under the appropriate assumptions to $HSA(\log K)$. □

[1] In ((Liu et al., 2016)), the Kernelized Stein discrepancy has been defined as KSD^2 and in (Gorham & Mackey, 2015) as KSD, hence it can be used interchangeably. For ease of the analysis, we proceed by considering KSD^2 as the Stein discrepancy.

G. Proof of Theorem 4.4

Proof. Consider the expression in (14) and we note that the Bayesian regret depends on the Stein information ratio $\mathbb{E}[\Gamma^*]$ and the total Stein information gain $\sum_{k=1}^{K}\mathbb{E}[\mathbb{K}_k^{\pi}\left(M^*;\mathcal{H}_{k,H}\right)]$. From Lemma 4.3, we can upper bound the right hand side of (14) as

$$\begin{aligned}\mathfrak{BR}_K \leq& \sqrt{SAH^3\cdot K\cdot HS^2A(\log K)} &(67)\\ =& H^2\sqrt{S^3A^2K\log K}=\tilde{\mathcal{O}}(H^2\sqrt{S^3A^2K}), &(68)\end{aligned}$$

where $\tilde{\mathcal{O}}$ absorbs the log factors and we obtain the final result. □

H. Proof of Theorem 4.6

Proof. Following the inequality $2ab\leq a^2+b^2$, for any policy π, we can write

$$\frac{\mathcal{R}_k}{\sqrt{\lambda\mathbb{K}_k^{\pi}\left(M^*;\mathcal{H}_{k,H}\right)}}\sqrt{\lambda\mathbb{K}_k^{\pi}\left(M^*;\mathcal{H}_{k,H}\right)}\leq\frac{\mathcal{R}_k^2}{2\lambda\mathbb{K}_k^{\pi}\left(M^*;\mathcal{H}_{k,H}\right)}+\frac{\lambda}{2}\mathbb{K}_k^{\pi}\left(M^*;\mathcal{H}_{k,H}\right), \tag{69}$$

where $\mathcal{R}_k:=\mathbb{E}\left[V_{1,\pi^*}^{M^*}(s_1^k)-V_{1,\pi^k}^{M^*}(s_1^k)\right]$,. Now, recollecting the definition of Bayesian regret from (3) and after adding and subtracting the regularization term, we can write

$$\mathfrak{BR}_K=\mathbb{E}\left[\sum_{k=1}^{K}\mathcal{R}_k-\lambda\sum_{k=1}^{K}\mathbb{K}_k^{\pi}\left(M^*;\mathcal{H}_{k,H}\right))+\lambda\sum_{k=1}^{K}\mathbb{K}_k^{\pi}\left(M^*;\mathcal{H}_{k,H}\right))\right]. \tag{70}$$

Utilizing the upper bound in (69), we can write

$$\begin{aligned}\mathfrak{BR}_K\leq&\frac{1}{2\lambda}\sum_{k=1}^{K}\mathbb{E}\left[\frac{\left(\mathbb{E}_k\left[V_{1,\pi^*}^{M^*}(s_1^k)-V_{1,\pi}^{M^*}(s_1^k)\right]\right)^2}{\mathbb{K}_k^{\pi}\left(M^*;\mathcal{H}_{k,H}\right))}\right]+\lambda\sum_{k=1}^{K}\mathbb{E}[\mathbb{K}_k^{\pi}\left(M^*;\mathcal{H}_{k,H}\right)]\\ =&\frac{K\mathbb{E}[\Gamma^*]}{2\lambda}+\lambda\sum_{k=1}^{K}\mathbb{E}[\mathbb{K}_k^{\pi}\left(M^*;\mathcal{H}_{k,H}\right)]. &(71)\end{aligned}$$

Now, select λ in (71) as $\lambda=\sqrt{K\mathbb{E}[\Gamma^*]/\sum_{k=1}^{K}\mathbb{E}[\mathbb{K}_k^{\pi}\left(M^*;\mathcal{H}_{k,H}\right)}$ to obtain the final expression. □

I. Detailed Information of Experimental Setup

I.1. Description of the Environments

DeepSea Environment: The DeepSea exploration environment (a slightly modified version of Osband & Van Roy (2017a) as used in Markou & Rasmussen (2019)) is an extremely challenging environment (see Fig. 8) to test the agent's capability of directed and sustained exploration. As shown in Fig. 8, there are total N states, the agent starts from the left-most state and can swim left or right from each of the N states in the environment. The agent gets a reward of $r=0$ (red transitions) for the left action. On the other hand, the right action from $s=1,\cdots,(N-1)$ succeeds with probability $(1-1/N)$, moving the agent to the right and otherwise fails and moving the agent to the left (blue arrows), giving $r\sim\mathcal{N}(-\delta,\delta^2)$ regardless of whether it succeeds. A successful swim-right from $s=N$ moves the agent back to $s=1$ and gives $r=1$. Hence, as we increase the number of states N, it will increase the amount of sparsity in the environment, making it extremely hard for the agent to explore (we provided this evidence in Fig. 2).

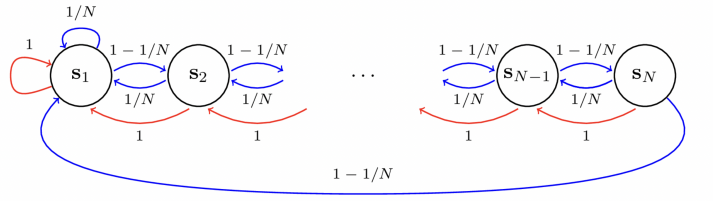


Figure 8. DeepSea Exploration Environment (Osband & Van Roy, 2017a; Markou & Rasmussen, 2019). This environment tests the agent's capability of sustained and directed exploration. This figure is same as Fig. 4 and we repeat it here for quick reference.

WideNarrow MDP Environment: This is another challenging environment (see Fig. 9) presented in Markou & Rasmussen (2019), which has $2N + 1$ states with deterministic transitions. In WideNarrow MDP environment, odd-numbered states except $s = (2N + 1)$ have W actions, out of which one gives $r \sim \mathcal{N}(\mu_l, \sigma_l^2)$ whereas all other actions result in $r \sim \mathcal{N}(\mu_h, \sigma_h^2)$, with $\mu_l < \mu_h$. Even-numbered states have a single action which results in $r \sim \mathcal{N}(\mu_h, \sigma_h^2)$. In the experiments, we use $mu_h = 0.5, \mu_l = 0, \sigma_l = \sigma_h = 1$. The WideNarrow MDP helps understand and compare the performance of `STEERING` under factored posterior approximations.

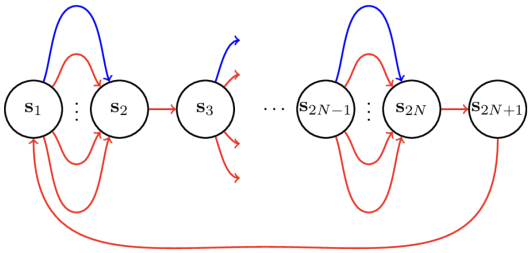


Figure 9. WideNarrow MDP environment (Markou & Rasmussen, 2019). The purpose of this environment is to test the agent's performance under factored posterior approximations.

PriorMDP Environment: The previous environments, DeepSea Exploration and WideNarrow MDP, have a special structure within them that tests different aspects of the agent/algorithm. In contrast, the PriorMDP environment of Markou & Rasmussen (2019) provides a more general environmental setup to test the algorithms where the dynamics are sampled from a Dirichlet prior with concentration $k = 1$ and reward from Normal prior with mean and precision drew from a Normal-Gamma prior with parameters $\langle 0, 1, 1, 4 \rangle$.

I.2. Baselines and Evaluations

We compare our proposed algorithm `STEERING` with various Bayesian/ non-Bayesian and distributional RL baselines. The baselines include

- **Q-learning**: Vanilla Q-learning with $\epsilon-$greedy action selection (Watkins & Dayan, 1992),
- **BQL**: Bayesian Q-learning (Dearden et al., 1998),
- **UBE**: Uncertainty Bellman equation (O'Donoghue et al., 2017),
- **MM**: Moment Matching across Bellman equation (Markou & Rasmussen, 2019),
- **PSRL**: Posterior Sampling Reinforcement Learning (Osband et al., 2013).

1. Bayesian Q-Learning : BQL introduced in Dearden et al. (1998) models the distribution over state-action returns Z^*, which is assumed to follow Gaussian (ergodic MDP) and updates the posterior belief of $P(\theta_{Z^*}|D)$ using Bayesian update rule. BQL considers Normal-Gamma prior on the parameters (mean and precision) of the Gaussian. However, there is a factored posterior assumption that restricts its generalisability and hinders the performance, as shown in Fig. 13

2. Uncertainty Bellman Equation: UBE proposed in O'Donoghue et al. (2017) is a model-based reinforcement learning approach designed primarily for modeling the epistemic uncertainty in $\mu_z = \mathbb{E}_z[Z|\theta_z]$ but with a strong assumption of MDP being a directed acyclic graph with bounded mean rewards. In other words, each state-action can be visited at most once per episode, which is restrictive and requires sparse design (repeating state-action multiple times) to make it work even for toy problems. Under these assumptions and a suitable Bellman operator, it holds that UBE has a unique solution which upper-bounds the epistemic uncertainty $\text{Var}_{\theta_T,\theta_R}[\mu_z]$ where θ_T, θ_R are the parameters of the transition and rewards model, respectively. However, the strong assumptions restrict the generalisability of the UBE-based methods to various practical scenarios where the inherent structure of the MDP is not acyclic.

3. Moment Matching across Bellman Equation: An interesting approach proposed recently by Markou & Rasmussen (2019) also uses the Bellman equation to estimate the epistemic uncertainty by comparing the moments (first and second moments) across the Bellman equation. While the first-order moments give the standard $V(s)$ & $Q(s,a)$, the second-order moments can be decomposed into aleatoric and epistemic terms without the need to compute upper bounds as in UBE-based methods. Similar to prior methods, the policy is optimized w.r.t $P(\theta_T|D)$, $P(\theta_R|D)$ and for the epistemic uncertainty μ_z. However, as with existing methods, MM also approximates with a factored posterior leading to performance loss.

4. Posterior Sampling Reinforcement Learning: Posterior Sampling reinforcement learning (PSRL) introduced by Osband et al. (2013); Osband & Van Roy (2014) is primarily built upon Thompson sampling or probability matching principle. PSRL provides an efficient and tractable solution to model-based RL with provable guarantees. The algorithm works by sampling a transition and rewards model from the posterior distribution at any k^{th} episode $\theta_T^k \sim P(\theta_T|D_k), \theta_R^k \sim P(\theta_R|D_k)$ and the optimal policy π^k is obtained by solving the Bellman equation under the sampled transition, rewards model. The agent then follows the policy π^k to interact with the environment and gather data and update the dictionary D_k. The primary advantage of PSRL lies in its computational tractability, as the policy needs to be optimized under a single sampled transition and rewards model. For PSRL, state-of-the-art Bayesian regret bounds exist under minor assumptions (Osband & Van Roy, 2014; 2017b). However, even the performance of PSRL degrades for complex environments such as under sparse reward scenarios, as empirically proved in Fig. 2(c).

We evaluate all the algorithms by comparing their performance in terms of cumulative regret. Further, for the Bayesian methods, we also evaluate the algorithms in terms of the posterior representation and concentration on the true Q^* values.

I.3. Implementation Details of `STEERING`

Here we present the implementation details of `STEERING`, as outlined in Algorithm 1. Our approach begins by sampling transition and reward models from the posterior distribution using Categorical-Dirichlet and Normal-Gamma distributions for the transition and rewards model, respectively, as in Markou & Rasmussen (2019). To ensure a fair comparison, we considered the same setup for all categorical and continuous distributions for the other baselines. In the second phase, our algorithm represents a distinct deviation from prior PSRL and IDS-based methods by quantifying the distributional distance between the true MDP and the current estimated MDP through the use of Stein discrepancy. Specifically, we utilize the regularized Stein information sampling as outlined in Theorem 4.6 for empirical analysis. Next, we compute the Stein-discrepency for each (s,a) pair using the formulae for DSD as defined in (13).

I.4. Hyperparameters

For all Dirichlet priors in the algorithms we use hyperparameters $\eta_{(s,a)} = 1$ and for Normal-Gamma priors we use $(\mu, \Lambda, \alpha, \beta)_{(s,a)} = (0, 4, 3, 3)$ as in Markou & Rasmussen (2019). For both `STEERING` and Var-IDS we use the same regularization constant (IA) $\lambda = 0.5$. For all the environments, we run the algorithms for $T = 5000$ timesteps with a buffer length (max) of size N, where N denotes the number of states and run the policy iteration for $2N$ iterations. We have also run PSRL, Var-IDS and `STEERING` for $T = 10000$ in Figure 1 to observe the effect of prior with convergence. For the baseline implementation of the algorithms including Q-Learning, UBE, BQL, MM, PSRL we leverage [2]. We utilize [3] and [4] for the DSD computation. We thank the authors for the open-source repositories.

[2] `https://github.com/stratisMarkou/sample-efficient-bayesian-rl`

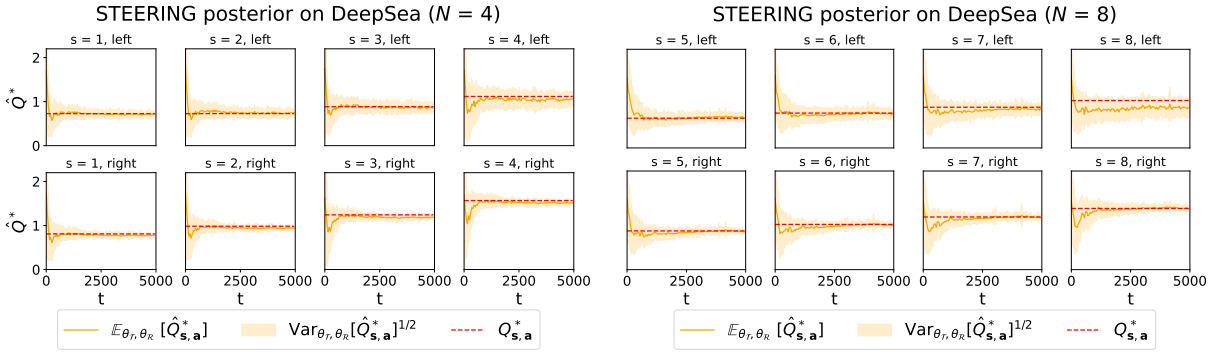
[3] `https://github.com/jiaseny/kdsd`

[4] `https://github.com/colehawkins/KSD-Thinning`

J. Additional Experimental Results & Discussions (Intuitive Insights)

J.1. Evolution of Posterior Representations for DeepSea Environment

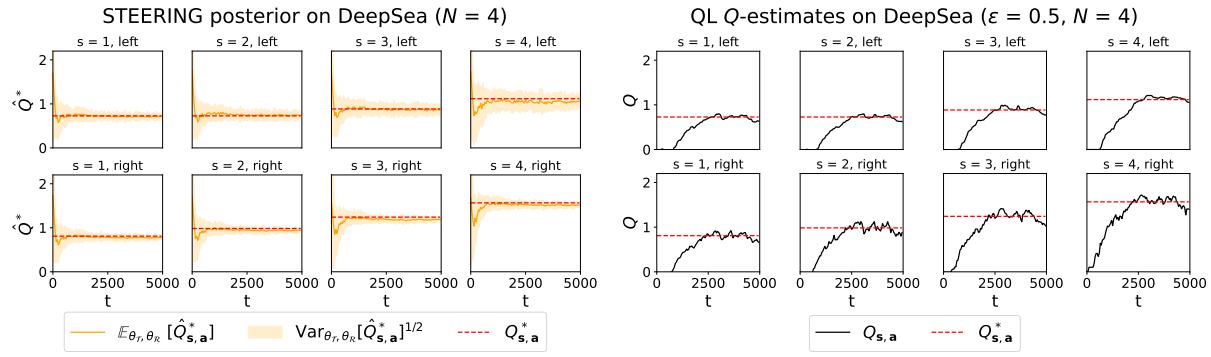
To gain more insights about the improvements and directed exploration behavior achieved by STEERING, we perform a detailed ablation study to analyze the evolution of posterior representation learned over iterations for all the algorithms. As a metric to plot, we plot the mean and variance of Q values calculated from the learned posterior and compare them against the true value denoted by Q^* (see Fig. 10 for STEERING). We remark that as the posterior concentrates with the progress in training, estimated Q values would also concentrates to the true Q^* (shown by dotted red line in Fig. 10). Since N denotes the number of states for DeepSea environment, and we obtain an estimated Q value for each state action pair, we choose to plot the evolution of the last 4 states -action pairs (left action in top row and right action in bottom row in Fig. 10).



(a) (b)

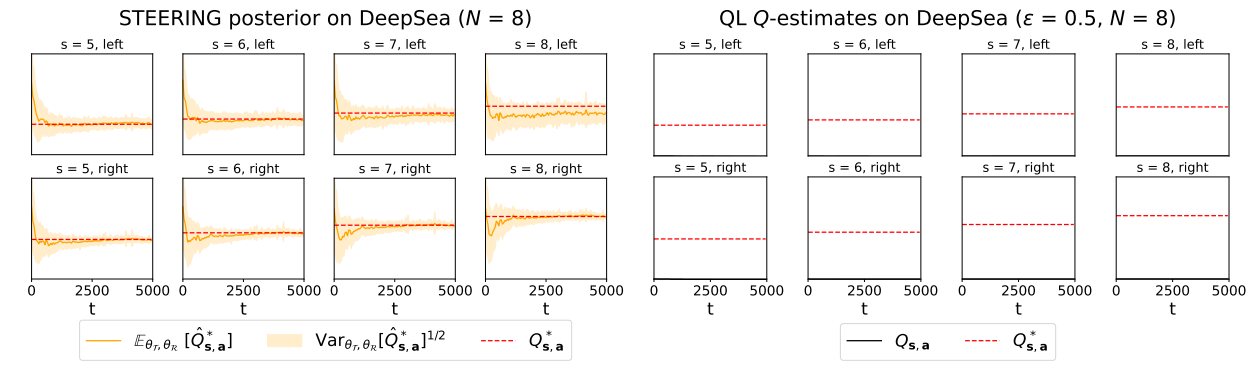
Figure 10. This plot analyzes the concentration of the predicted $\hat{Q}$ to the true Q^* (ground-truth) with iterations for STEERING. **(a)** This figure shows the performance for DeepSea environment with $N = 4$. **(b)** This figure shows the performance for DeepSea environment with $N = 8$.

Fig. 10 is of critical importance, as it gives a clear understanding of whether the agent is over-exploring or under-exploring actions based on its sub-optimality. To make the advantage clear as compared to existing approaches, we plot similar figures for all the existing algorithms in Fig. 11 to Fig. 19. **Interestingly, in all the comparison plots from Fig. 11 to Fig. 19, STEERING exhibits a superior posterior representation which results in directed exploration, even with increased sparsity.** The most interesting aspect of STEERING over baselines is that it does not over-explore actions once it is confident that those are sub-optimal which is an important feature helps in achieving directed exploration.



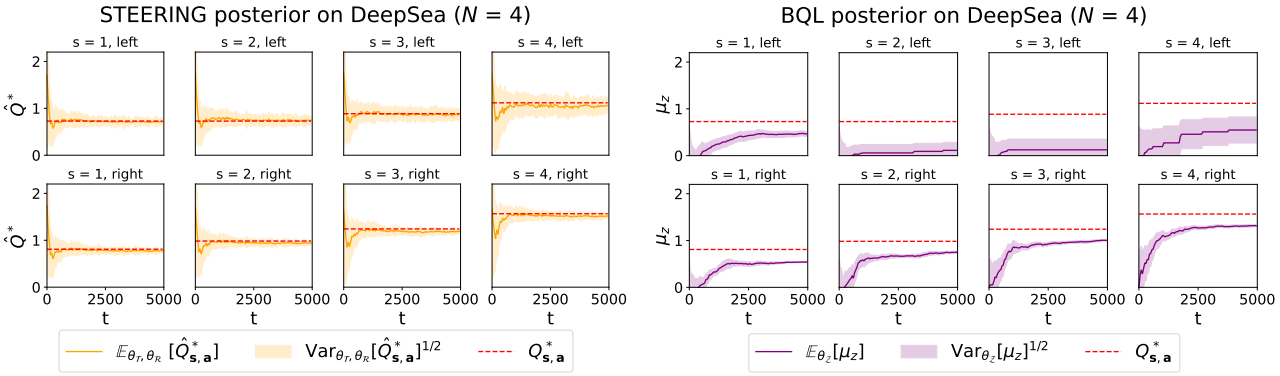
(a) (b)

Figure 11. Comparison of STEERING and Q-learning on DeepSea Exploration with sparsity $N = 4$. This plot shows the concentration of the predicted $\hat{Q}$ to the true Q^* (ground-truth) versus iterations. It is evident that STEERING converges to true Q^* much faster than Q-learning. **Remark:** As right actions are optimal for DeepSea environment, STEERING stops exploring left actions beyond a point, leading to a comparatively higher variance in $\hat{Q}$ for left actions, thus providing directed exploration.



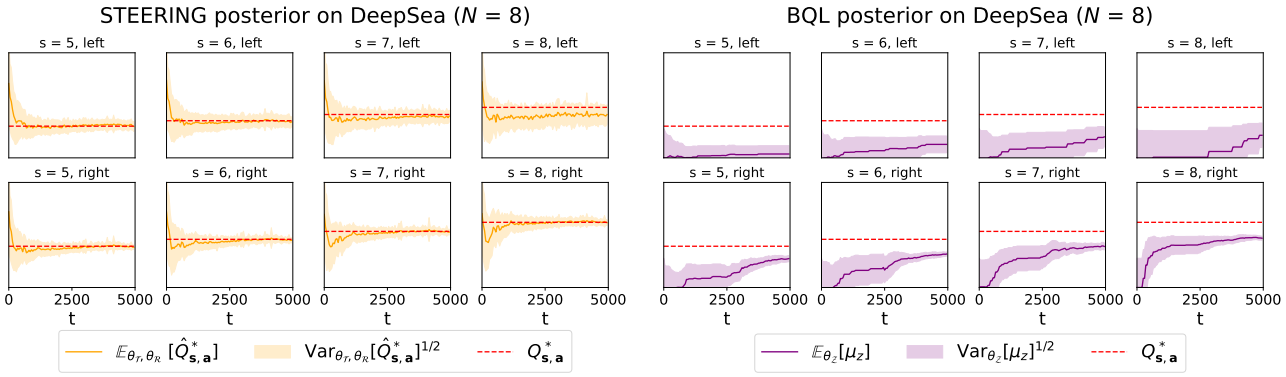
(a) (b)

Figure 12. Comparison of STEERING and Q-learning on DeepSea Exploration with sparsity $N = 8$. This plot shows the concentration of the predicted $\hat{Q}$ to the true Q^* (ground-truth) with iterations. **Remark:** As the sparsity is increased, the performance of Q-learning degrades drastically (potentially due to lack of efficient exploration) whereas STEERING converges to true Q^* efficiently. Further, we note that since right actions are optimal for DeepSea environment, STEERING stops exploring left actions beyond a point leading to a comparatively higher variance in $\hat{Q}$ for left actions, thus providing directed exploration.



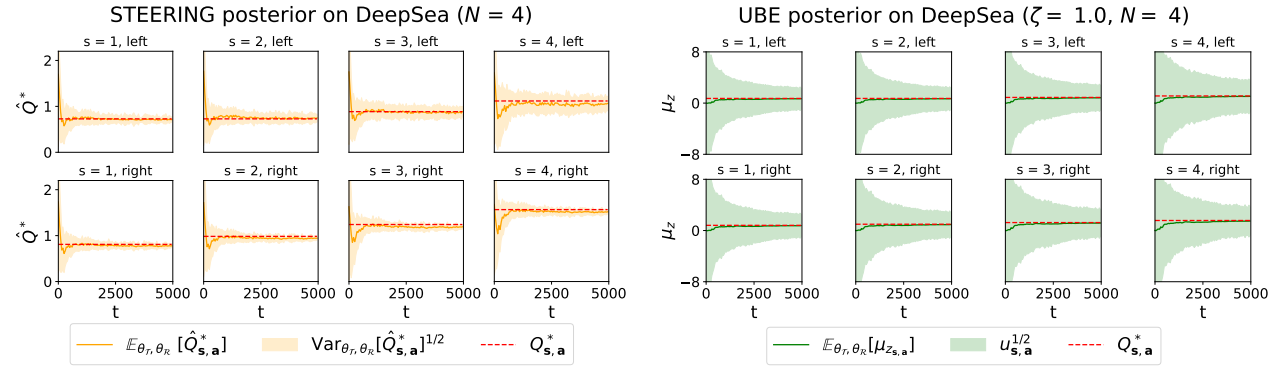
(a) (b)

Figure 13. Comparison of STEERING and Bayesian-Q learning (BQL) on DeepSea Exploration with sparsity $N = 4$. This plot shows the concentration of the predicted $\hat{Q}$ to the true Q^* (ground-truth) versus iterations. **Remark:** STEERING converges to true Q^* much faster than Bayesian-Q learning which fails to concentrate on the true Q^*. This is because BQL does not have an efficient forgetting mechanism in its update rule leading to high dependence on inaccurate past updates. This leads to the observation that the posterior is overconfident about incorrect predictions in BQL. However, STEERING stops exploring left actions beyond a point as right actions are optimal for DeepSea environment, leading to a comparatively higher variance in $\hat{Q}$ for left actions, but low variance for right actions, thus providing directed exploration.



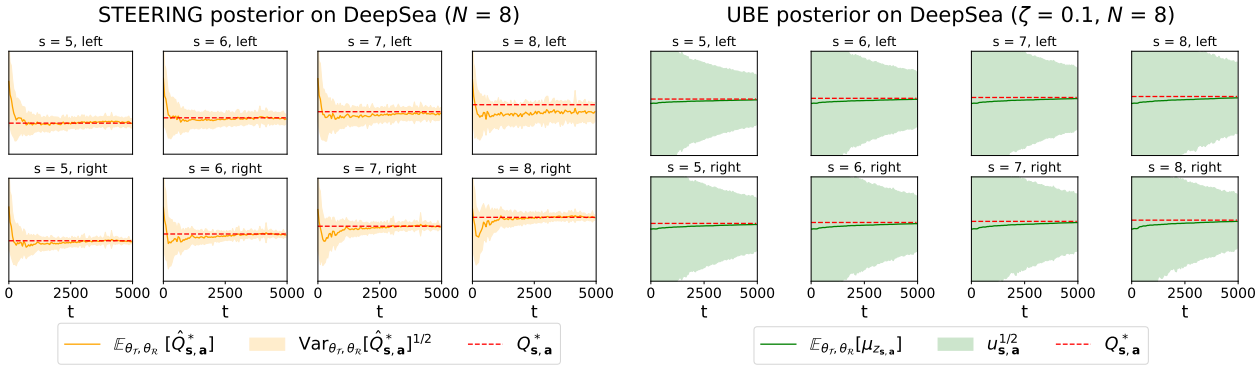
(a) (b)

Figure 14. Comparison of STEERING and Bayesian-Q learning (BQL) on DeepSea Exploration with sparsity $N = 8$. This plot shows the concentration of the predicted $\hat{Q}$ to the true Q^* (ground-truth) versus iteration with more sparsity. **Remark:** STEERING converges to true Q^* much faster than Bayesian-Q learning which fails to concentrate on the true Q^*. A major reason can be that BQL doesn't have an efficient forgetting mechanism in its update rule leading to high dependence on inaccurate past updates. Hence, we observe that the posterior for BQL is overconfident about incorrect predictions. In contrast, STEERING stops exploring left actions beyond a point as right actions are optimal for DeepSea Exploration environment, leading to a comparatively higher variance in $\hat{Q}$ for left actions, thus providing directed exploration.



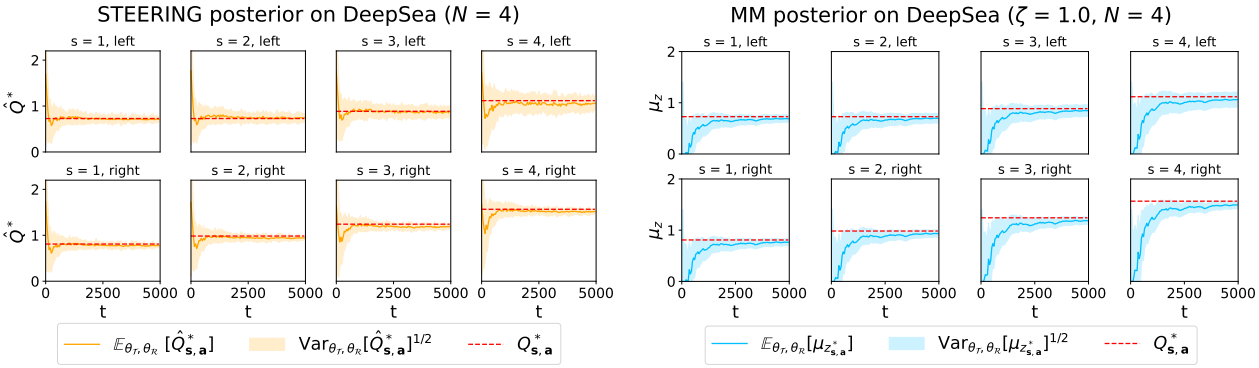
(a) (b)

Figure 15. Comparison of STEERING and uncertainty Bellman equation (UBE) on DeepSea Exploration with sparsity $N = 4$. This plot shows the concentration of the predicted $\hat{Q}$ to the true Q^* (ground-truth) versus iterations. **Remark:** STEERING converges much more efficiently to true Q^* (or μ_z^*) compared to UBE which fails to concentrate properly. For UBE, even if the predicted mean is closer to optima, the variance is too high which leads to sub-optimal exploration. In contrast, STEERING stops exploring left actions beyond a point as right actions are optimal for the DeepSea Exploration environment, leading to a comparatively higher variance in $\hat{Q}$ for left actions, thus providing directed exploration.



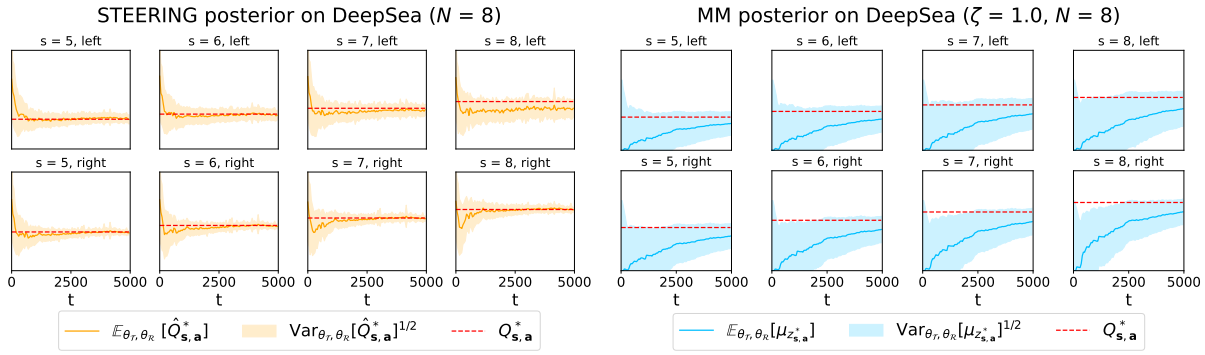
(a) (b)

Figure 16. Comparison of STEERING and uncertainty Bellman equation (UBE) on DeepSea Exploration with sparsity $N = 8$. This plot analyzes the concentration of the predicted $\hat{Q}$ to the true Q^* (ground-truth) versus iterations with more sparsity. **Remark:** As the sparsity is increased ($N = 8$), the variance in estimated μ_z^* for UBE increases significantly and the performance degrades drastically leading to random action selection. In contrast, STEERING converges much more efficiently to true Q^* (or μ_z^*). Also, STEERING stops exploring left actions beyond a point as right actions are optimal for DeepSea Exploration environment, leading to a comparatively higher variance in $\hat{Q}$ for left actions, thus providing directed exploration.



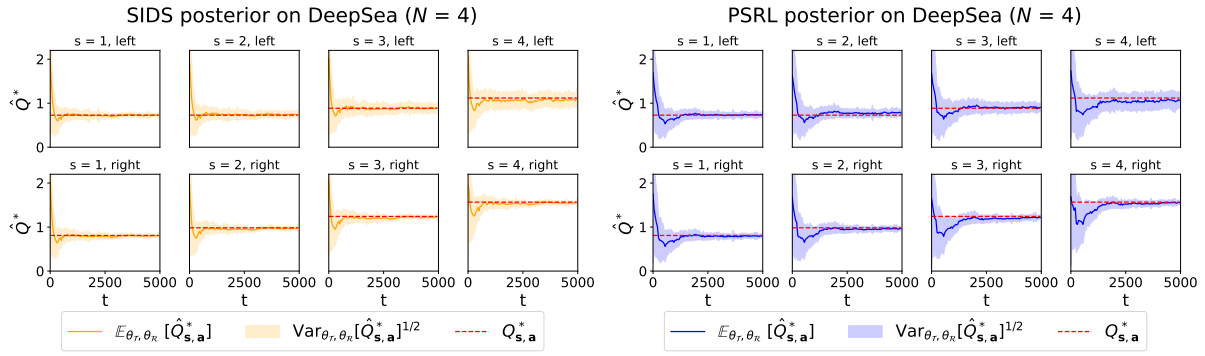
(a) (b)

Figure 17. Comparison of STEERING and moment matching (MM) on DeepSea Exploration with sparsity $N = 4$. This plot analyzes the concentration of the predicted $\hat{Q}$ to the true Q^* (or μ_z^*) versus iterations. **Remark:** Although MM performs well in this setting, STEERING converges more efficiently and faster to true Q^*. Also, STEERING unlike MM stops exploring left actions beyond a point as right actions are optimal for DeepSea Exploration environment, leading to a comparatively higher variance in $\hat{Q}$ for left actions and low variance for right actions, thus providing directed exploration.



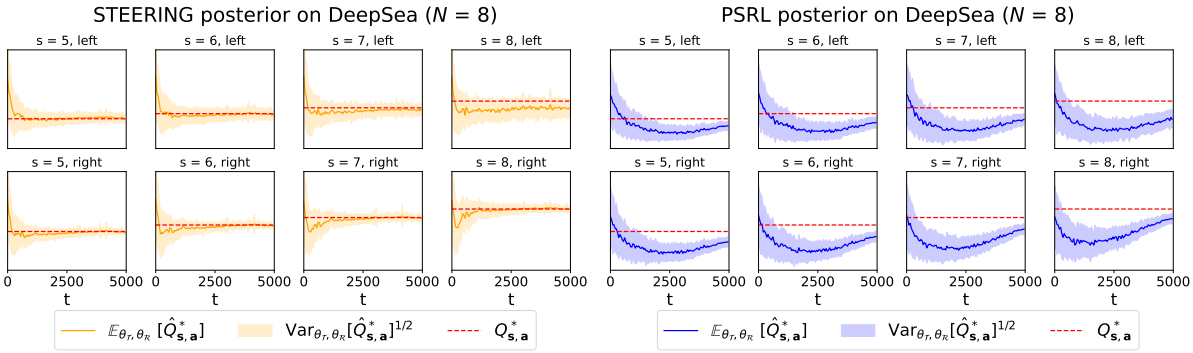
(a) (b)

Figure 18. Comparison of STEERING and moment matching on DeepSea Exploration with sparsity $N = 8$. This plot analyzes the concentration of the predicted $\hat{Q}$ to the true Q^* (or μ_z^*) (ground-truth) versus iterations. **Remark:** As the sparsity is increased ($N = 8$), the performance of MM degrades with increased variance of predicted μ_z and also converges to sub-optimal mu_z. While STEERING converges to true Q^* efficiently. Also, since right actions are optimal for DeepSea Exploration environment, STEERING stops exploring left actions beyond a point leading to a comparatively higher variance in $\hat{Q}$ for left actions, thus providing directed exploration.



(a) (b)

Figure 19. Comparison of STEERING and posterior sampling reinforcement learning (PSRL) on DeepSea Exploration with sparsity $N = 4$. This plot analyzes the concentration of the predicted $\hat{Q}$ to the true Q^* versus iterations. **Remark:** The performance of PSRL is comparable to STEERING, but still the convergence is faster with lesser variance for STEERING. We also note that the sparsity level for this setting is low ($N = 4$).



(a) (b)

Figure 20. Comparison of STEERING and posterior sampling reinforcement learning (PSRL) on DeepSea Exploration with sparsity $N = 8$. This plot analyzes the concentration of the predicted $\hat{Q}$ to the true Q^* (ground-truth) versus iterations. **Remark:** As the sparsity is increased ($N = 8$), the performance of PSRL degrades with slower convergence to Q^* with higher variance in $\hat{Q}$ prediction. In contrast, STEERING converges much more efficiently to true Q^*. Also, STEERING stops exploring left actions beyond a point as right actions are optimal for DeepSea Exploration environment, leading to a comparatively higher variance in $\hat{Q}$ for left actions, thus providing directed exploration.

J.2. Convergence Plots for DSD

Finally, we show the correctness of our approach by observing the convergence of DSD in Fig. 21. We note that for the optimal actions (orange), the DSD converges to much lower values than sub-optimal actions (blue) which in-turn implies the effectiveness of our Stein-information based proposed approach.

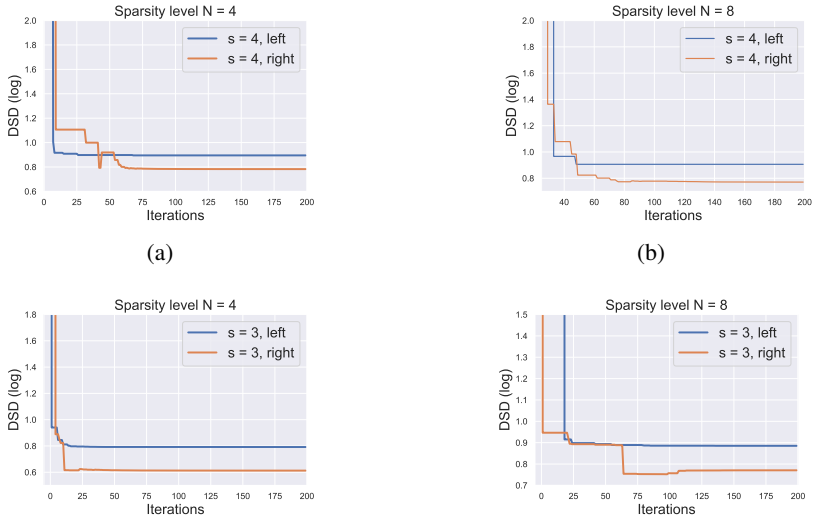


Figure 21. This figure provides evidence of DSD convergence (mean plot over 5 seeds) for different state-action pairs and sparsity levels with iterations for the DeepSea environment. Additionally, the plot also provides an indication of directed exploration through the DSD convergence to lower values for right actions which moves agent towards goal than left in states $s = 3$ and $s = 4$.

J.3. Additional Comparisons for WideNarrow MDP and PriorMDP

Here in Fig. 22, we also compares the performance of `STEERING` on WideNarrow MDP and PriorMDP environments (Markou & Rasmussen, 2019; Osband & Van Roy, 2017a) against the existing RL baselines : vanilla Q-learning with $\epsilon-$greedy action selection (Watkins & Dayan, 1992), Bayesian Q-learning (BQL) (Dearden et al., 1998), Uncertainty Bellman Equation (UBE) (O'Donoghue et al., 2017), Moment matching (MM) across Bellman equation (Markou & Rasmussen, 2019), Posterior Sampling RL (PSRL) (Osband et al., 2013), and IDS (Hao & Lattimore, 2022). We approximated information gain with variance to implement IDS, hence denoted by Var-IDS in Fig. 22.

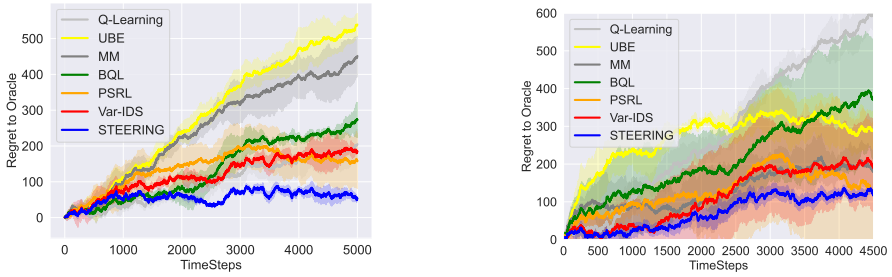


Figure 22. This figure compares the performance of `STEERING` on (a) WideNarrow MDP and (b) PriorMDP environments (Markou & Rasmussen, 2019; Osband & Van Roy, 2017a). **Remark:** WideNarrow MDP tests the algorithm's ability under factored posterior approximations, whereas PriorMDP tests the algorithm's ability to more general environments without specific structures. We note that `STEERING` outperforms existing baselines in both the environments.